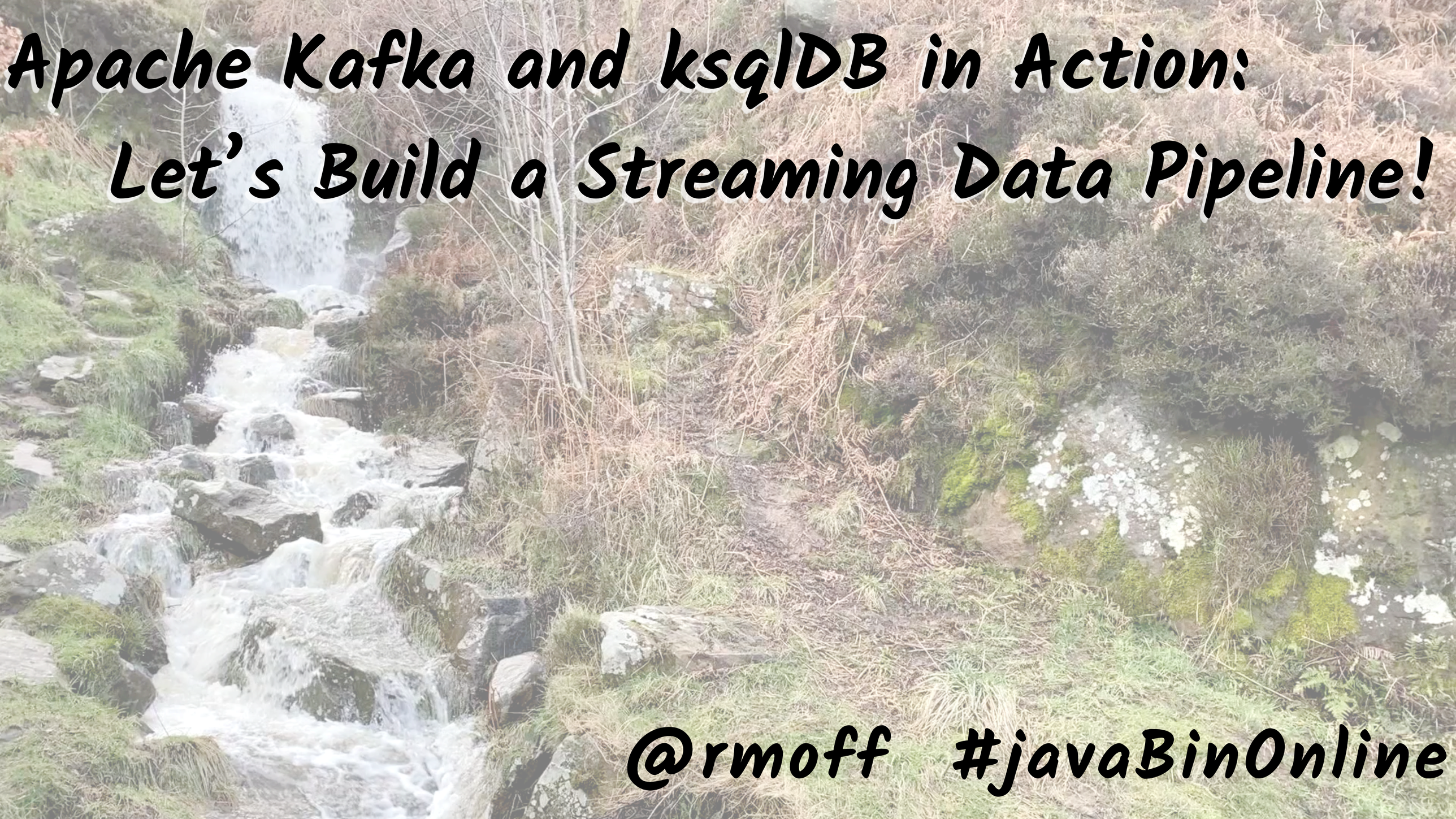
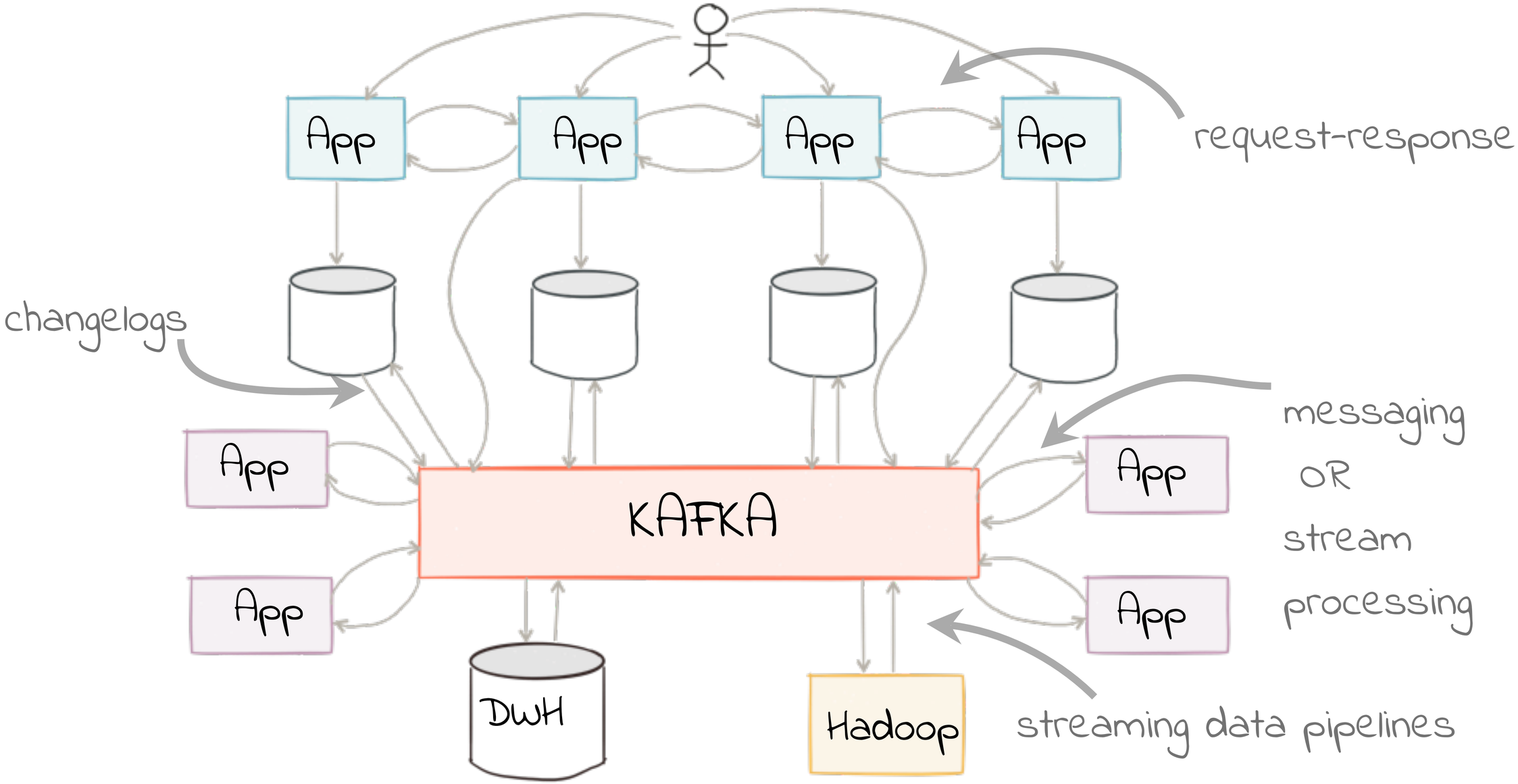
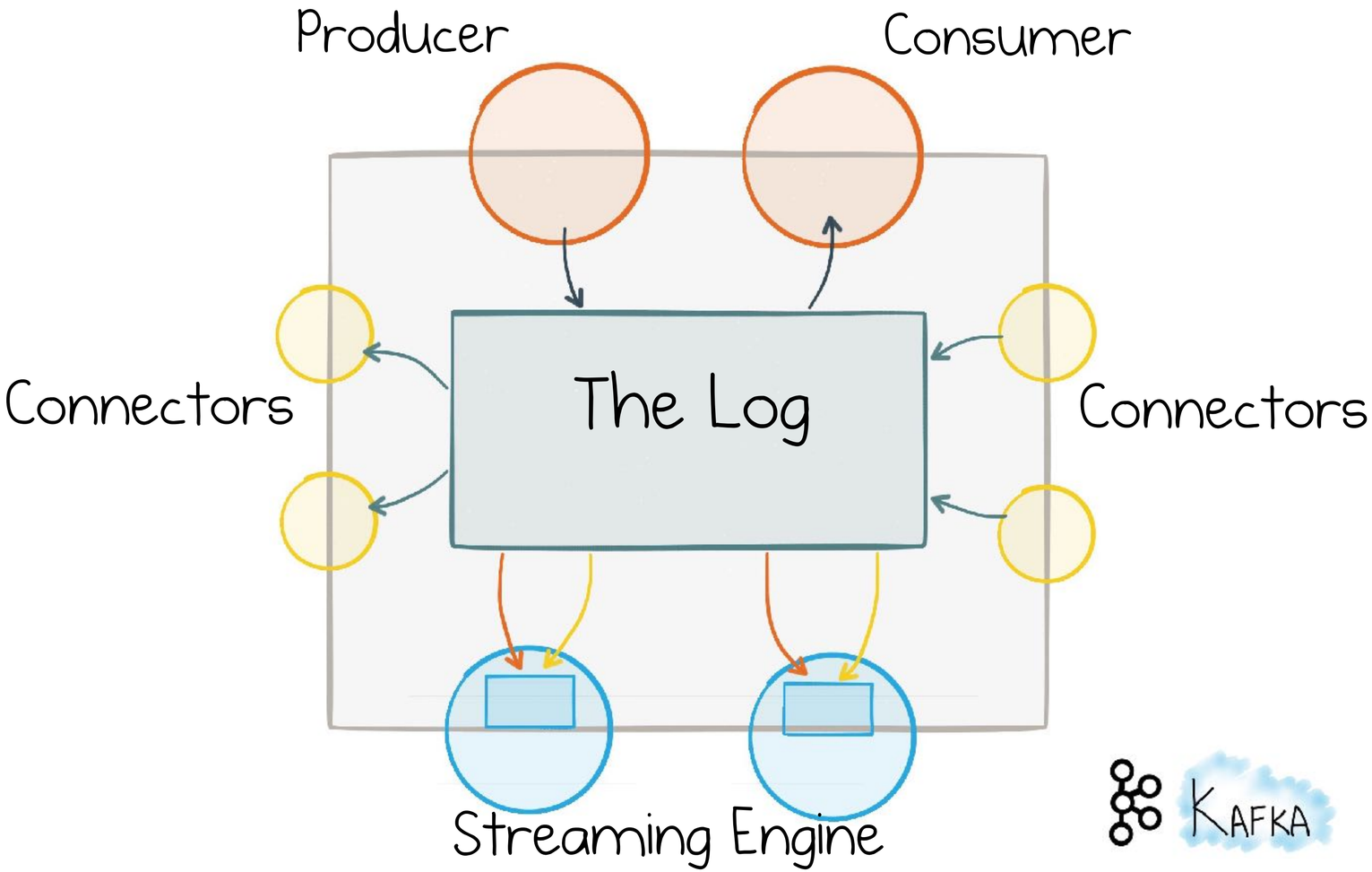

Apache Kafka and ksqlDB in Action:
Let's Build a Streaming Data Pipeline!
@rmoff #javaBinOnline

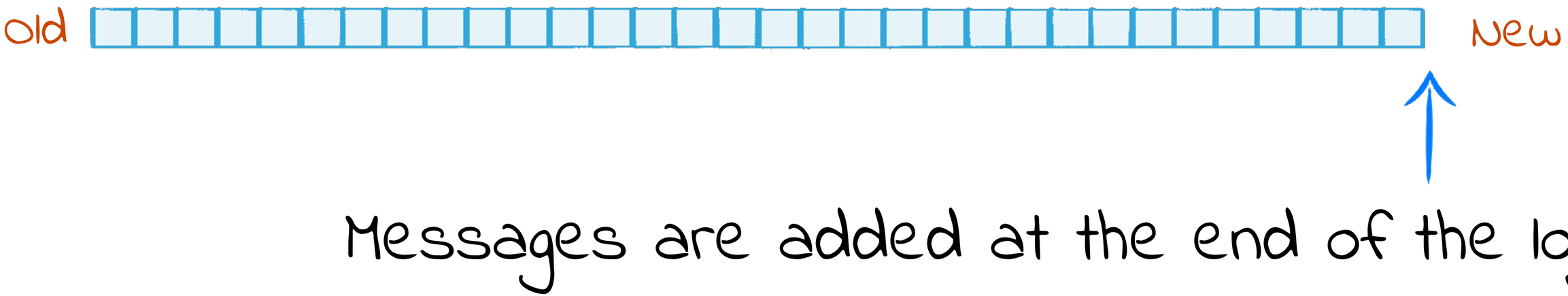
Kafka is an Event Streaming Platform



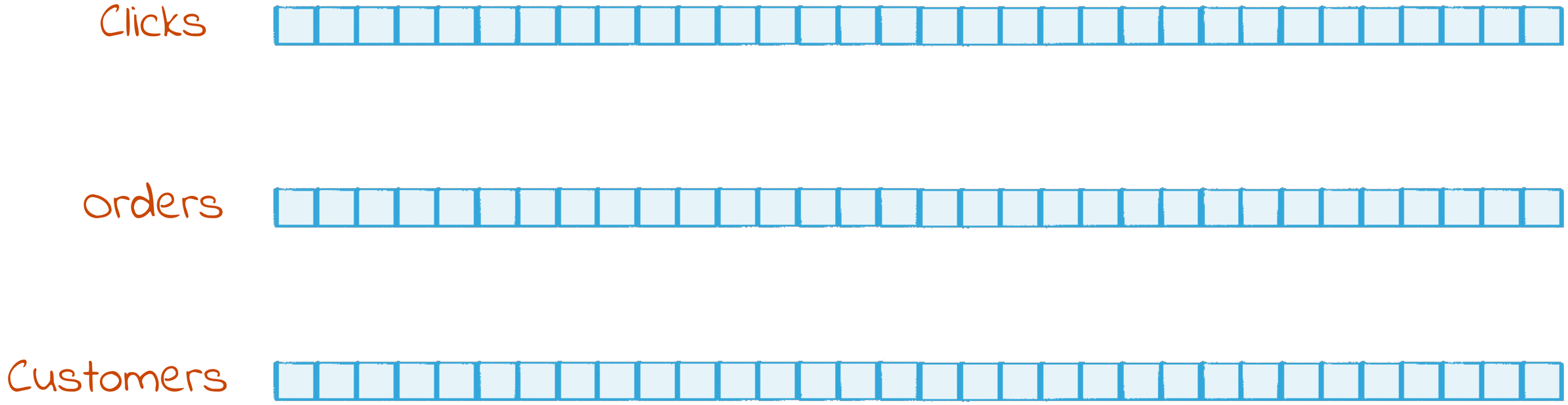
What is an Event Streaming Platform?



Immutable Event Log

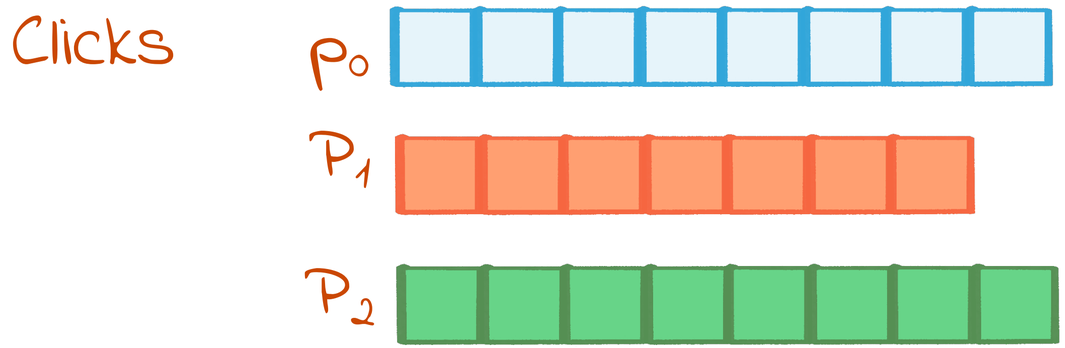


Topics



Topics are similar in concept to tables in a database

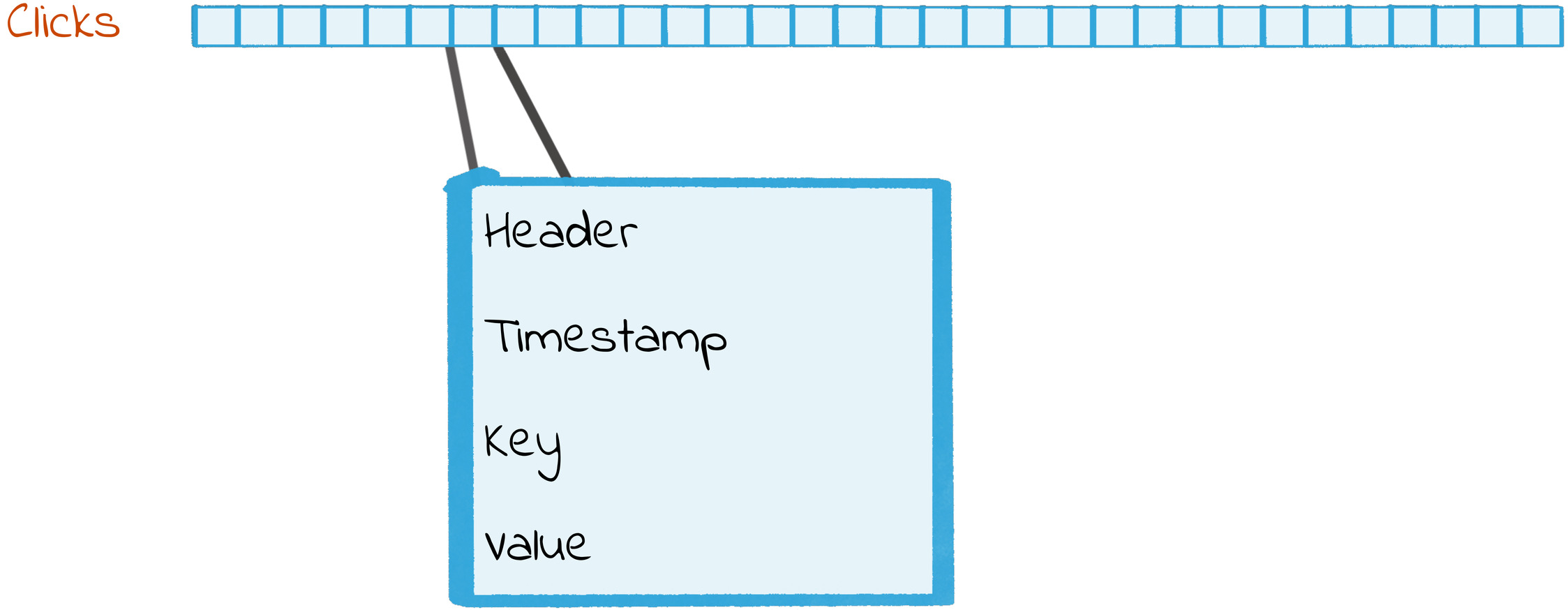
Partitions



Messages are guaranteed to be strictly ordered within a partition

Messages are just KN bytes

plus headers + timestamp



Messages are just KN bytes

With great power comes great responsibility

Avro

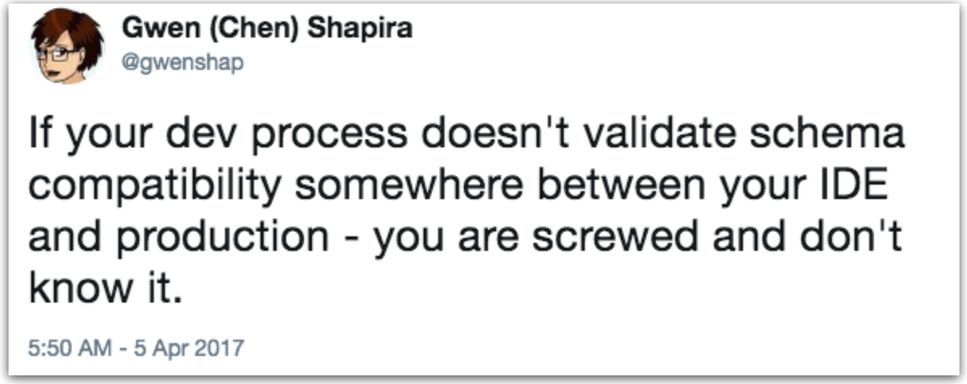
-> Confluent Schema Registry

Protobuf

-> Confluent Schema Registry

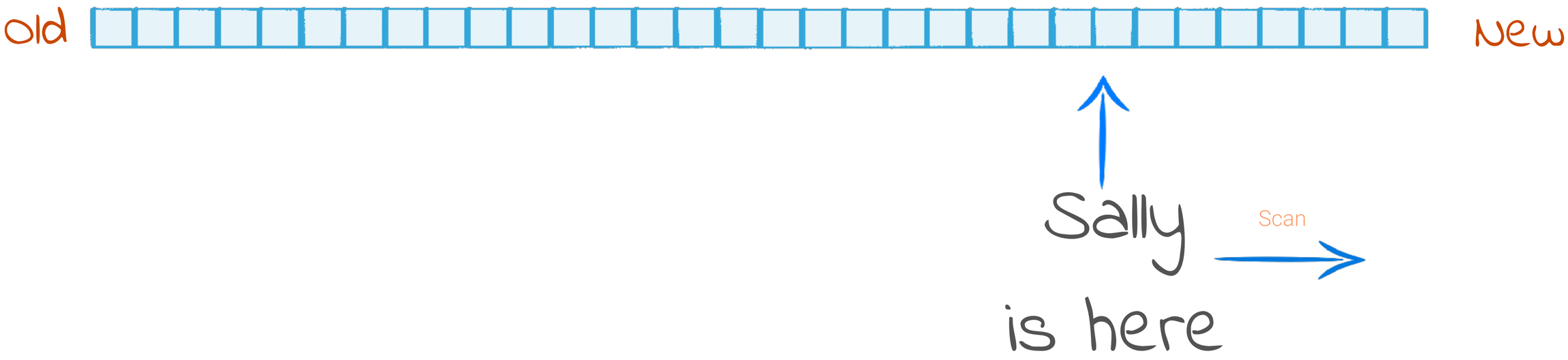
JSON

CSV

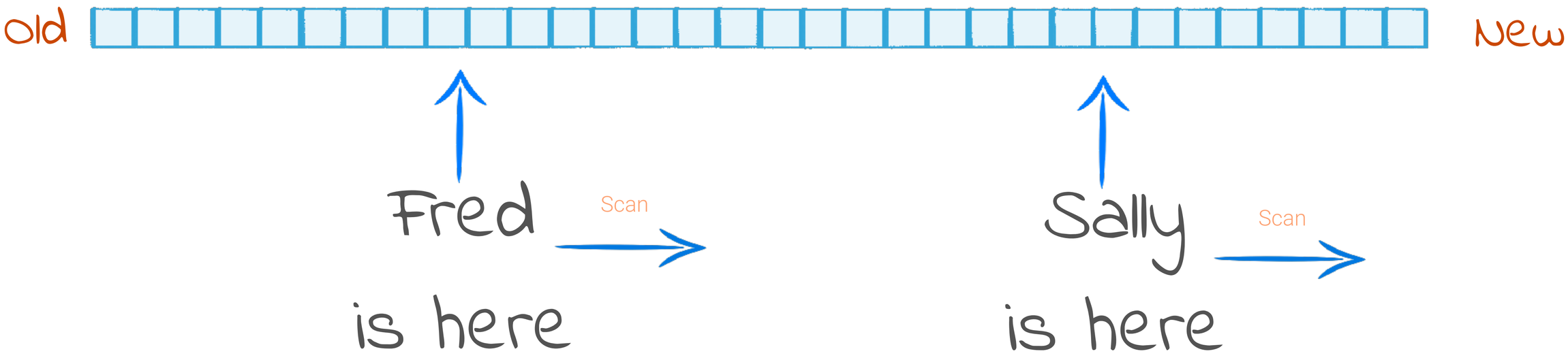


https://qconnewyork.com/system/files/presentation-slides/qcon_17_-_schemas_and_apis.pdf

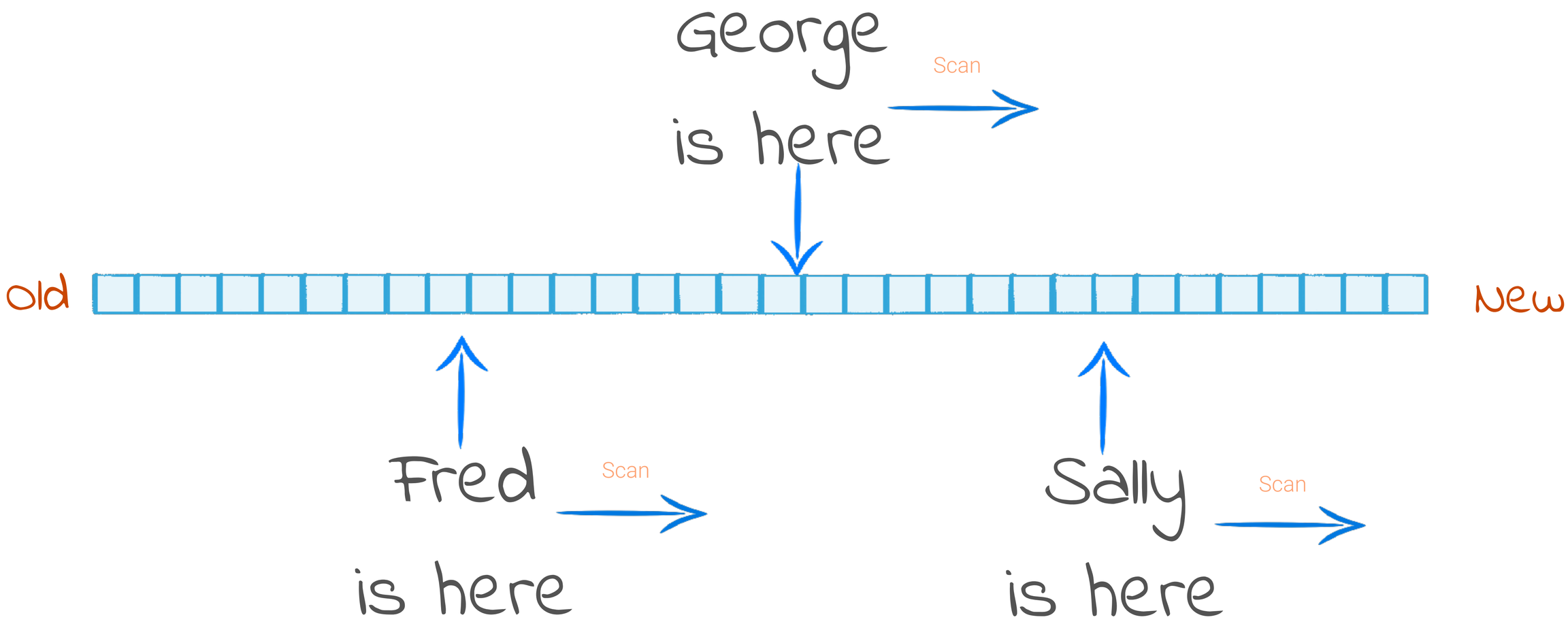
Consumers have a position all of their own



Consumers have a position all of their own

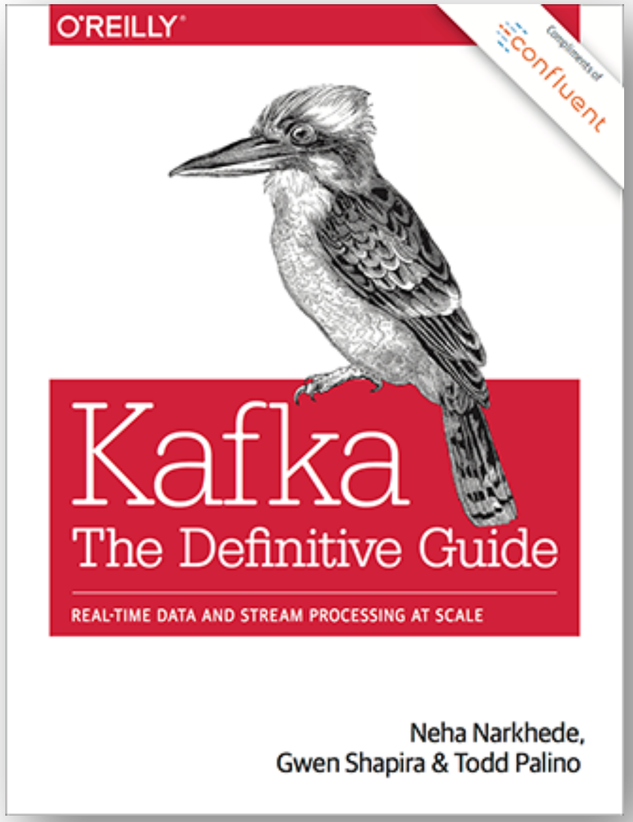


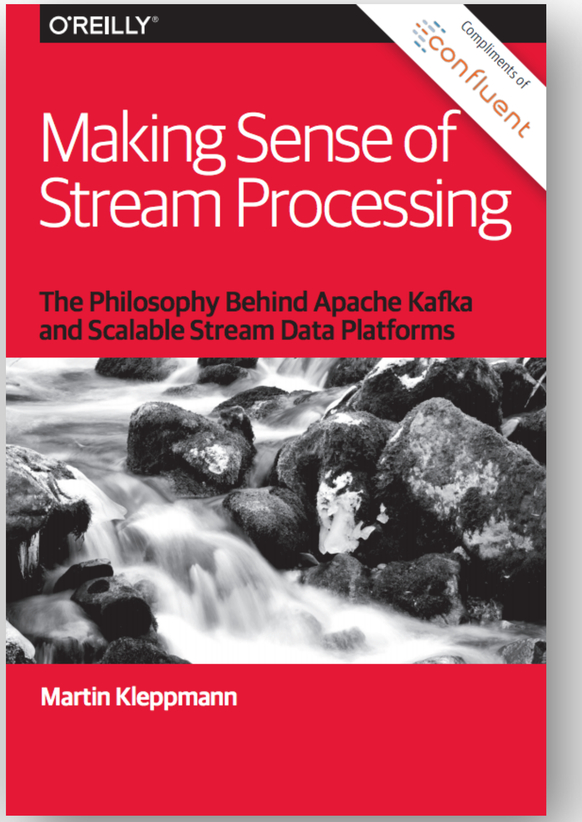
Consumers have a position all of their own

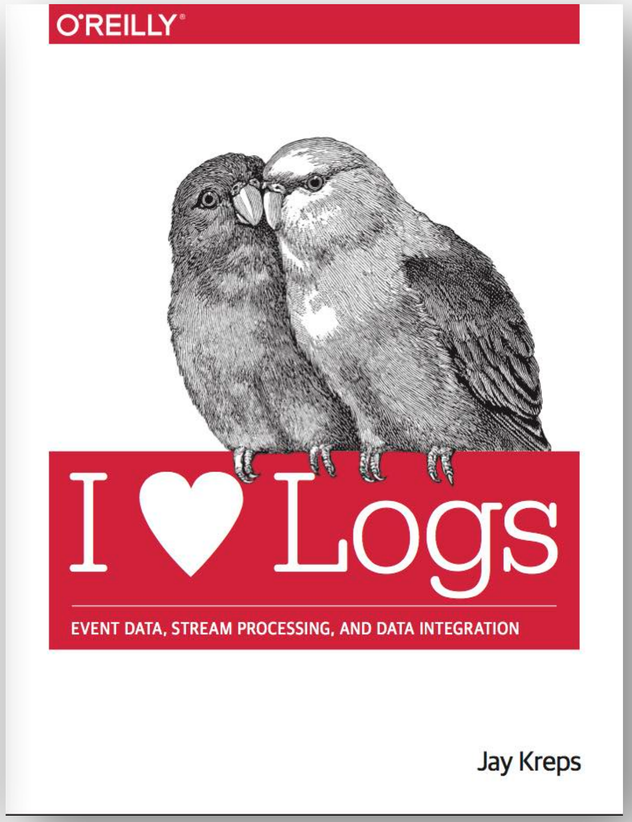


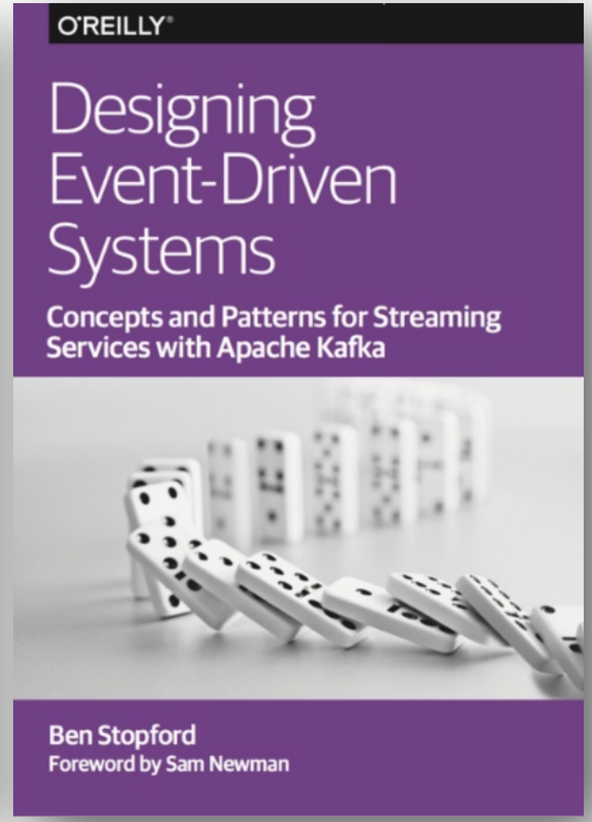
Free Books!

https://rmoff.dev/javabin







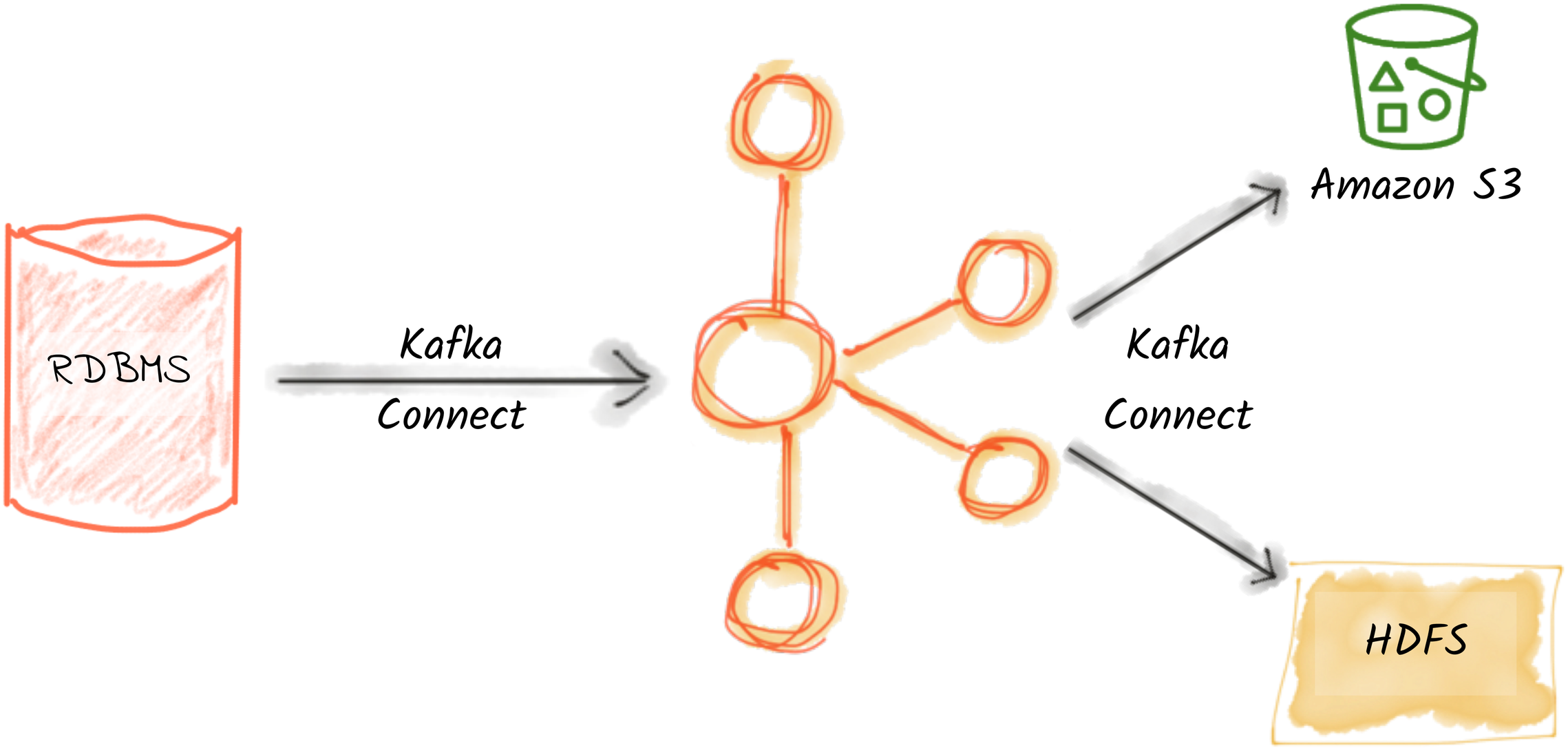


Streaming Data Pipelines

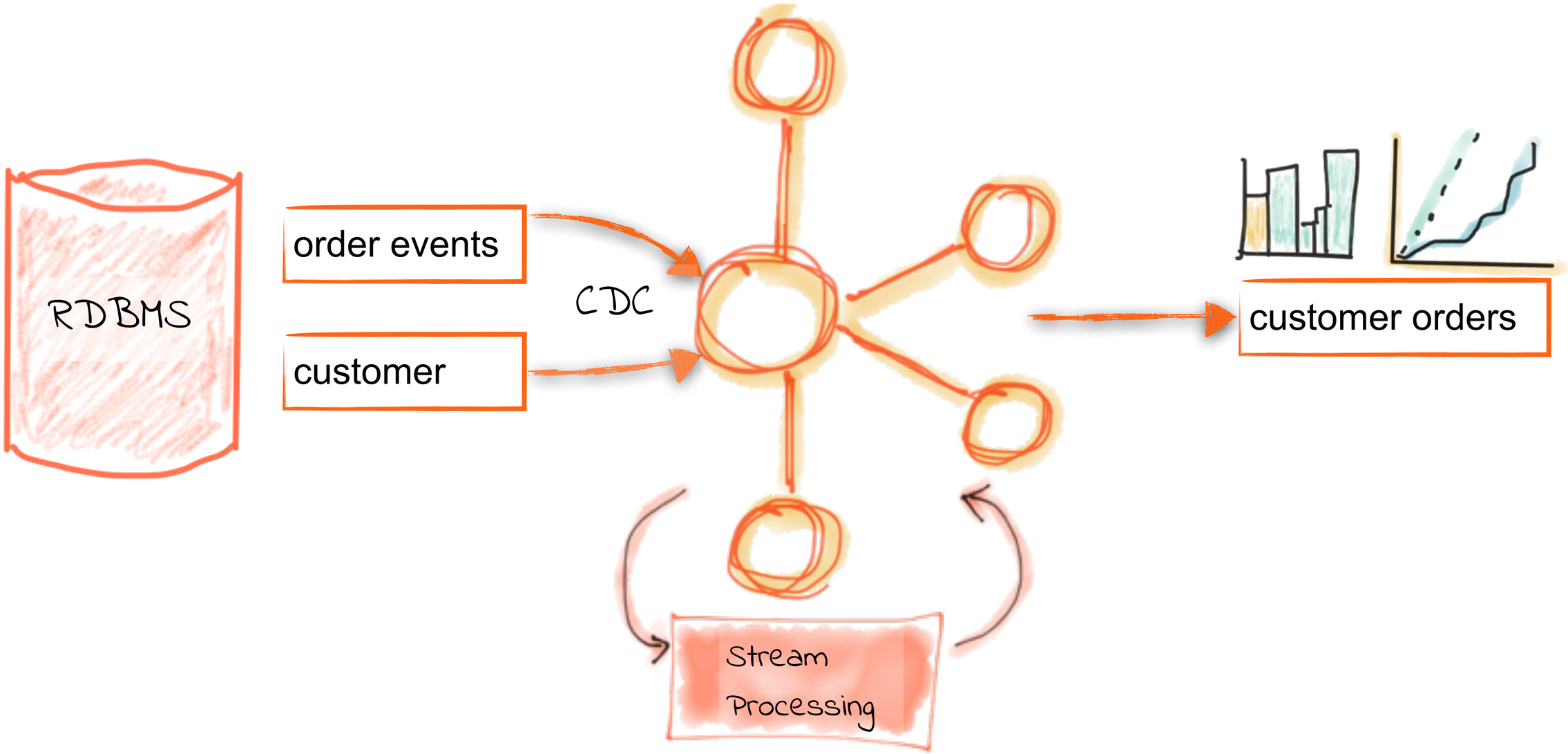
Photo by Victor Garcia on Unsplash

confluent

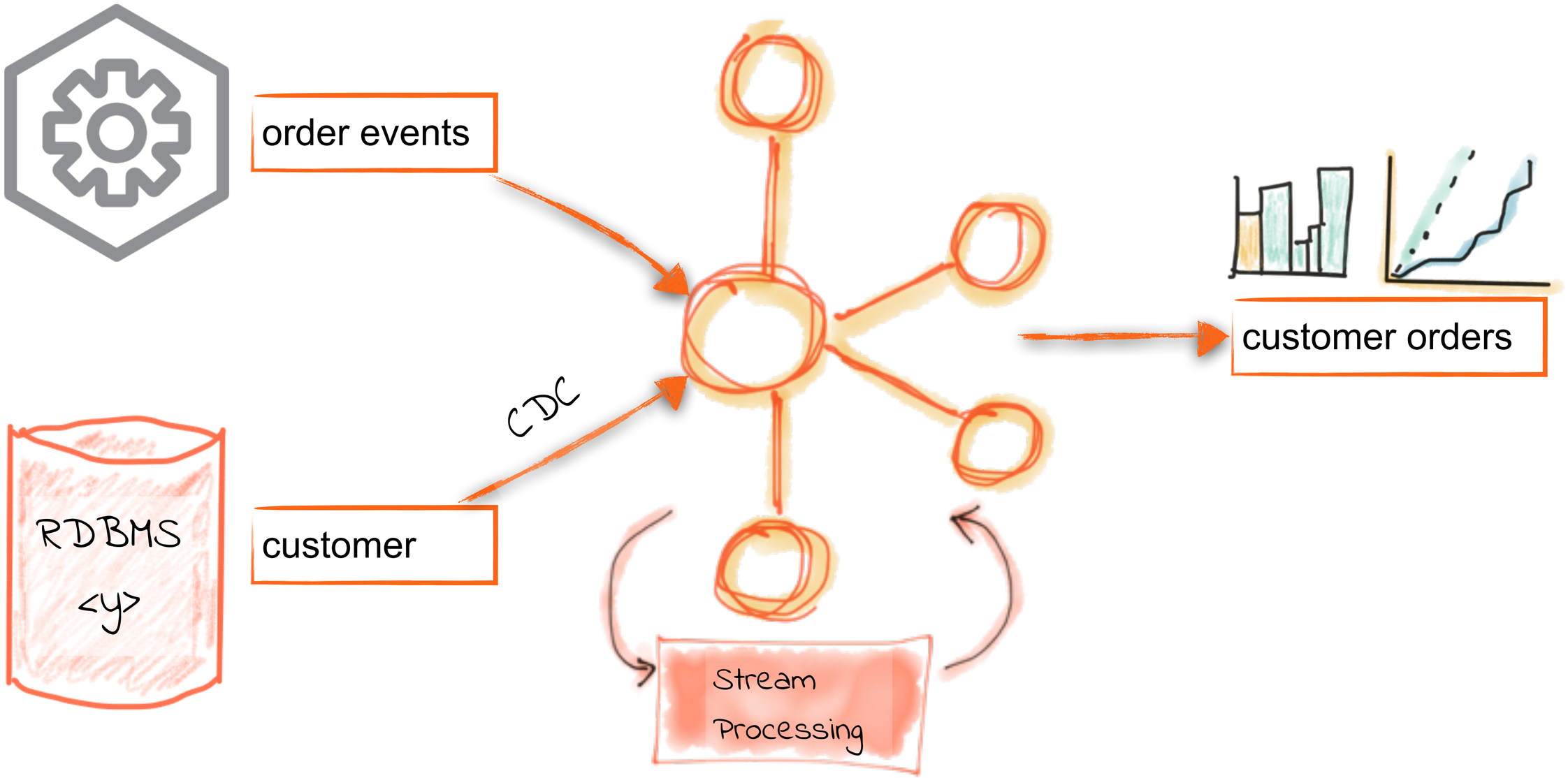
Database Offload



Stream Processing with Apache Kafka and ksqlDB



Real-time Event Stream Enrichment



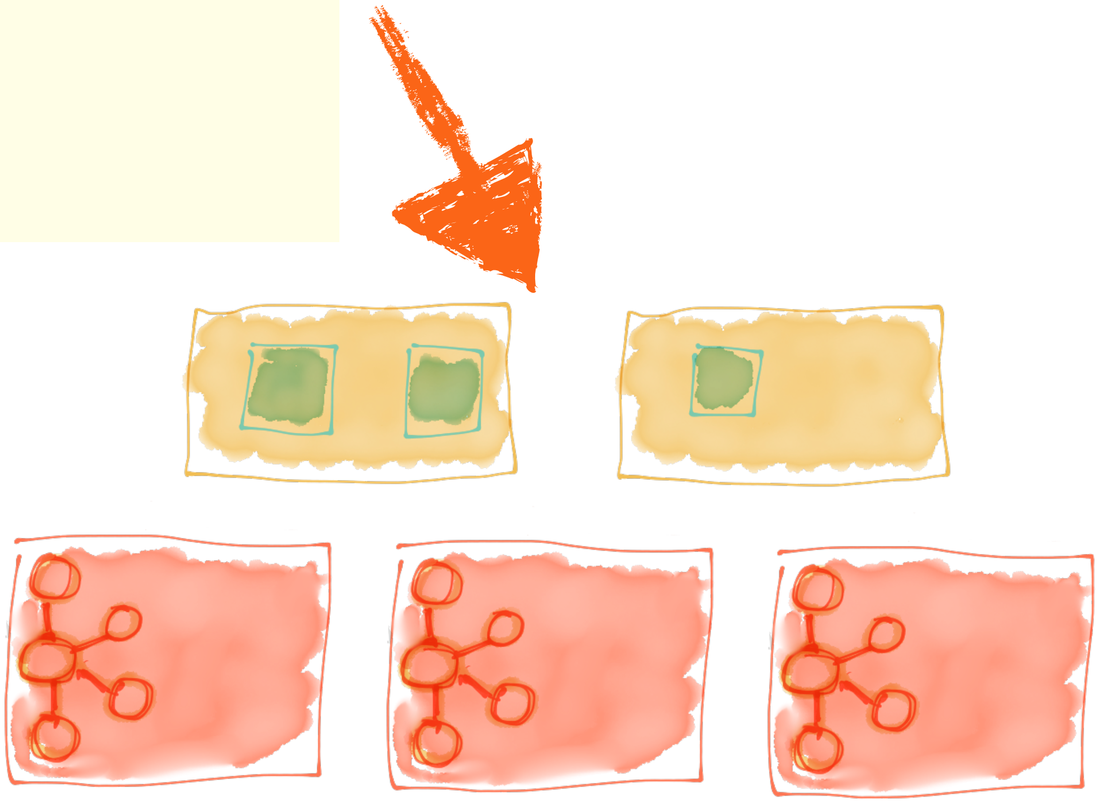
Building a streaming data pipeline

Integration + Stream Processing

Integration

Streaming Integration with Kafka Connect

Sources

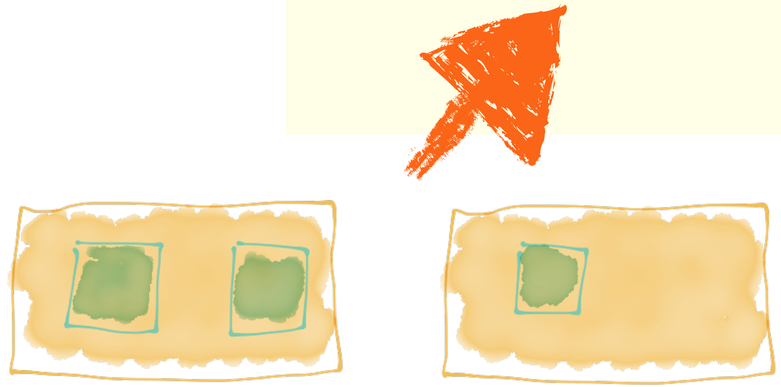


Kafka Connect

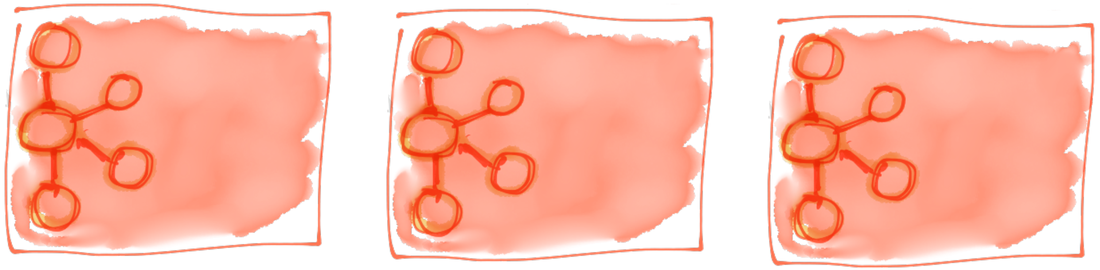
Kafka Brokers

Streaming Integration with Kafka Connect

Sinks

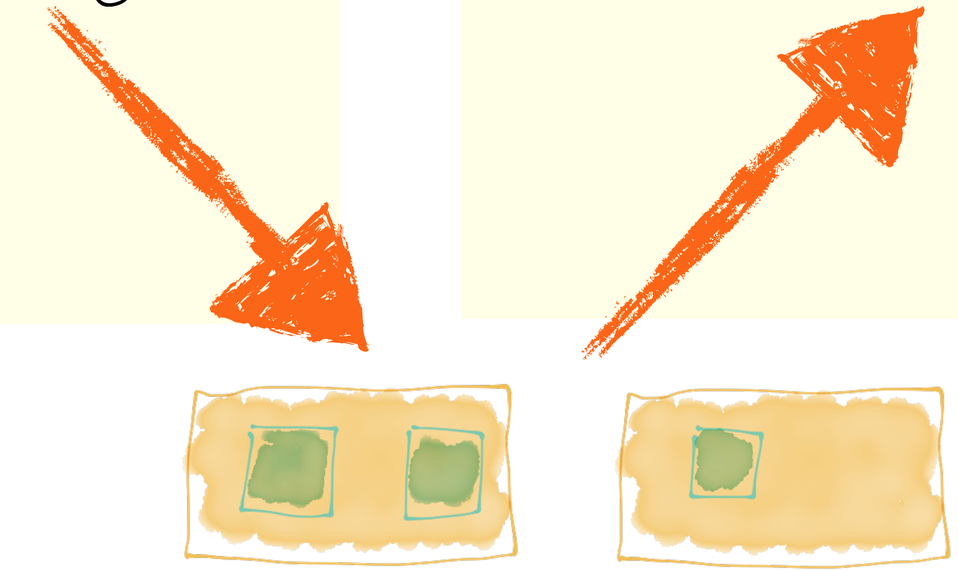


Kafka Connect

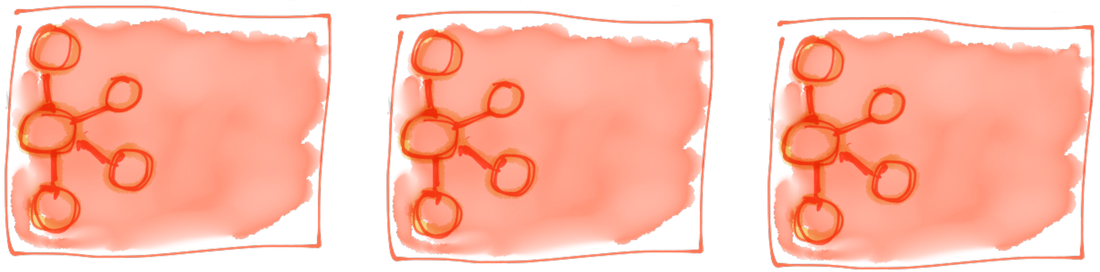


Kafka Brokers

Streaming Integration with Kafka Connect



Kafka Connect



Kafka Brokers

Look Ma, No Code!

```
{
  "connector.class":
    "io.confluent.connect.jdbc.JdbcSourceConnector",
  "connection.url":
    "jdbc:mysql://asgard:3306/demo",
  "table.whitelist":
    "sales,orders,customers"
}
```

https://docs.confluent.io/current/connect/

Messages are just KN bytes

With great power comes great responsibility

Avro

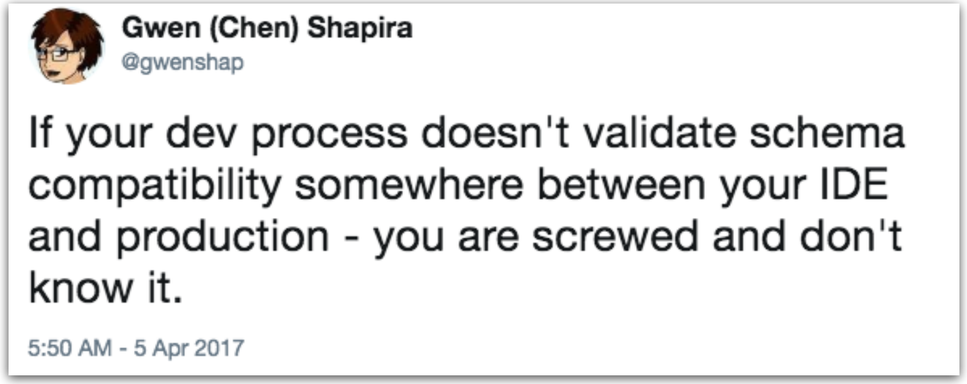
-> Confluent Schema Registry

Protobuf

-> Confluent Schema Registry

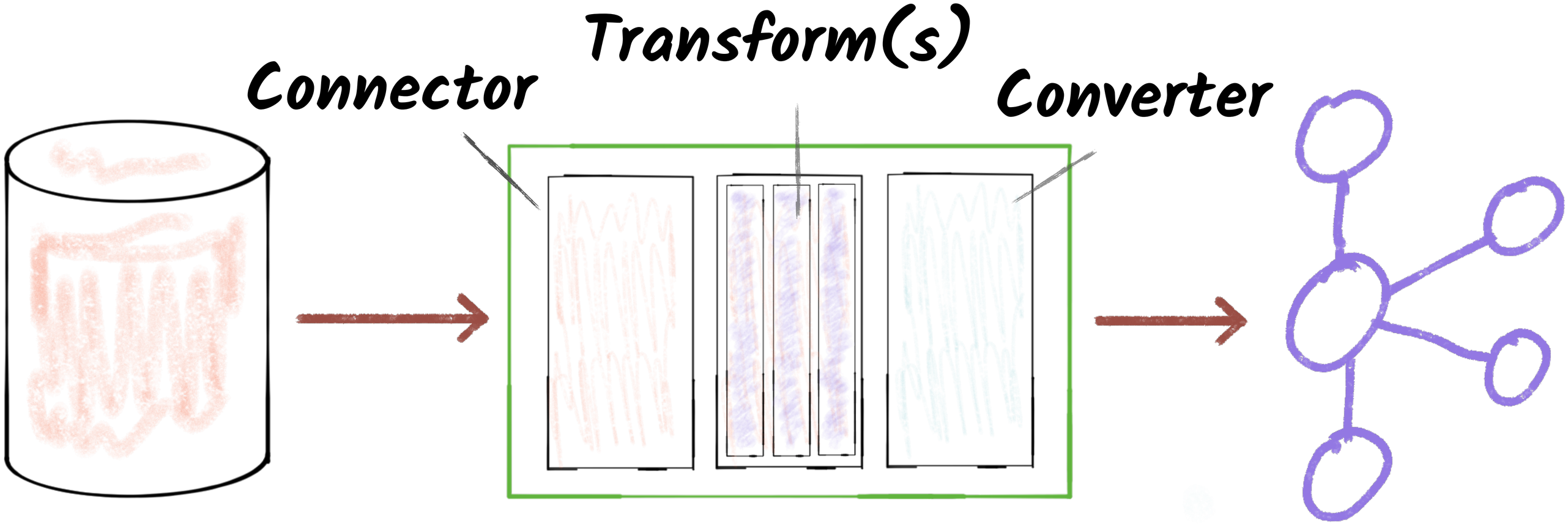
JSON

CSV

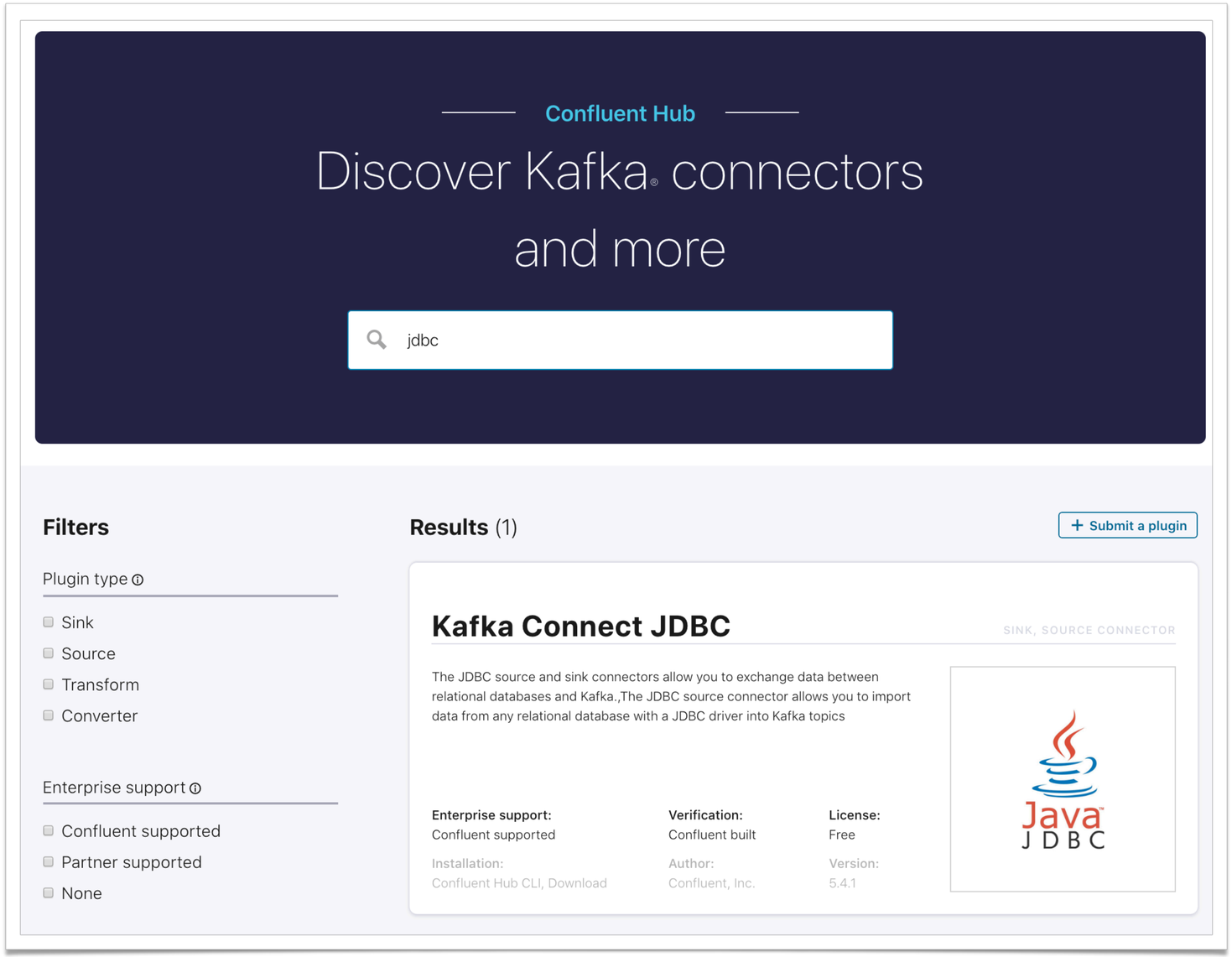


https://qconnewyork.com/system/files/presentation-slides/qcon_17_-_schemas_and_apis.pdf

Extensible



Confluent Hub



hub.confluent.io

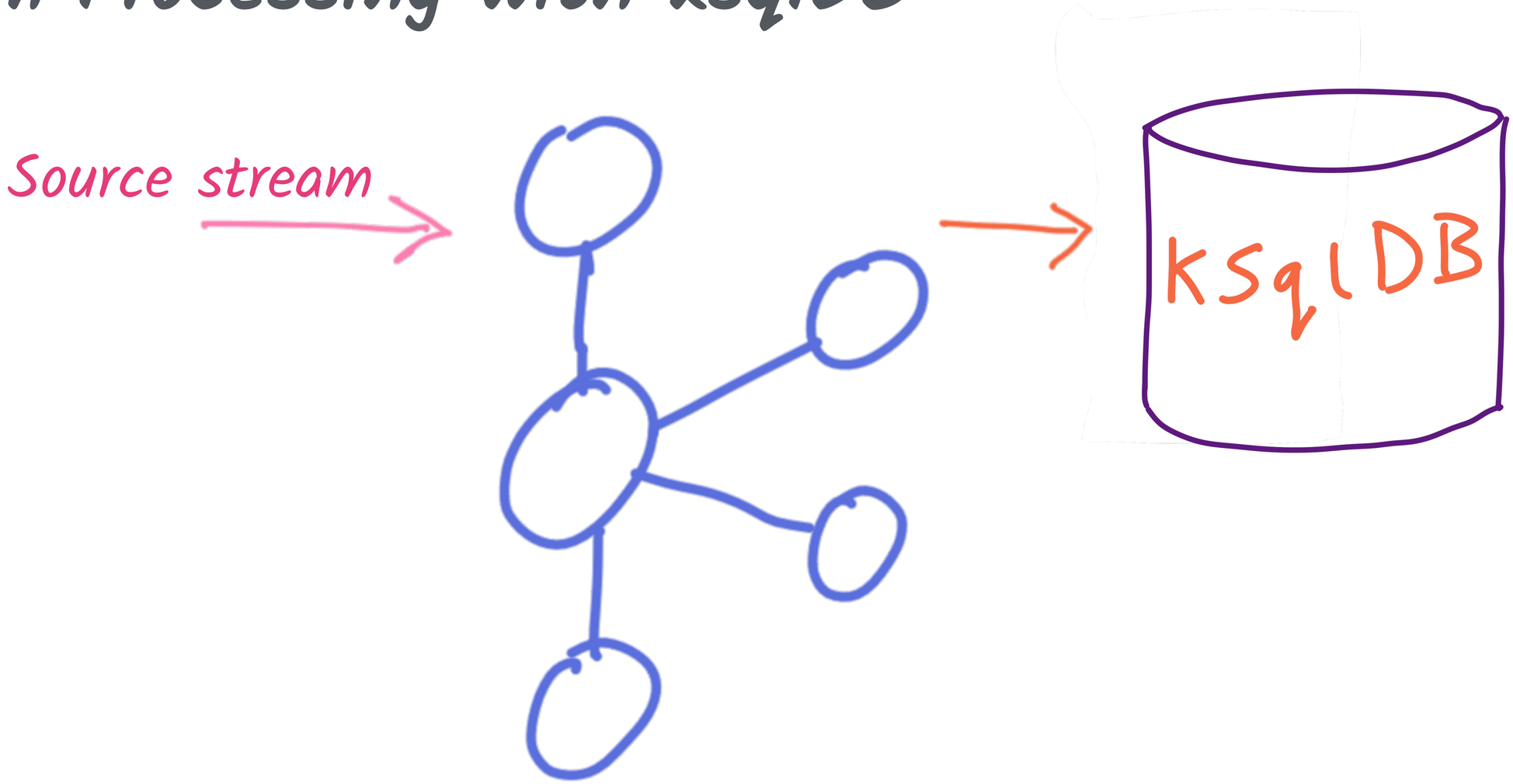
Integration + Stream Processing

Stream Processing

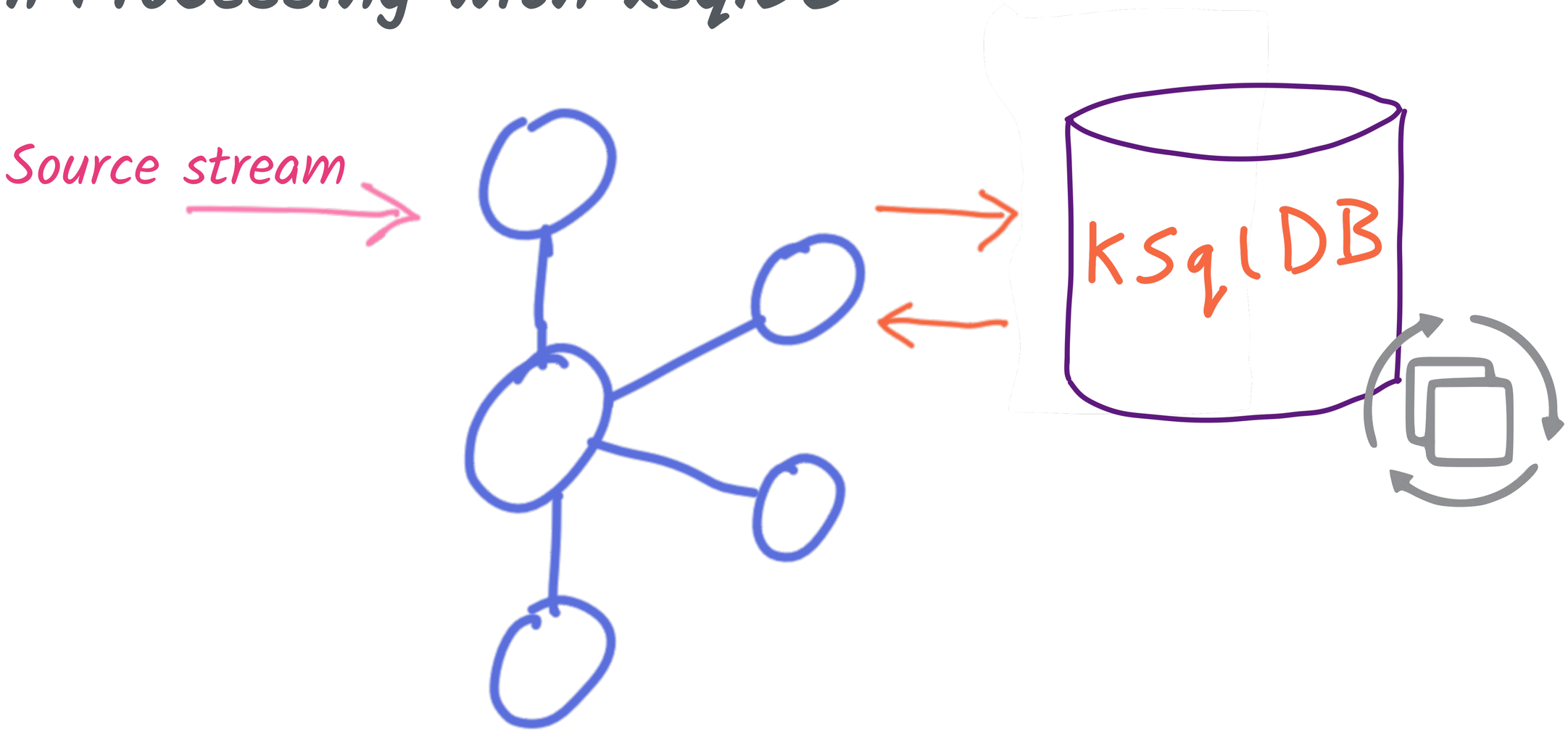
ksqlDB

The event streaming database purpose-built for stream processing applications.

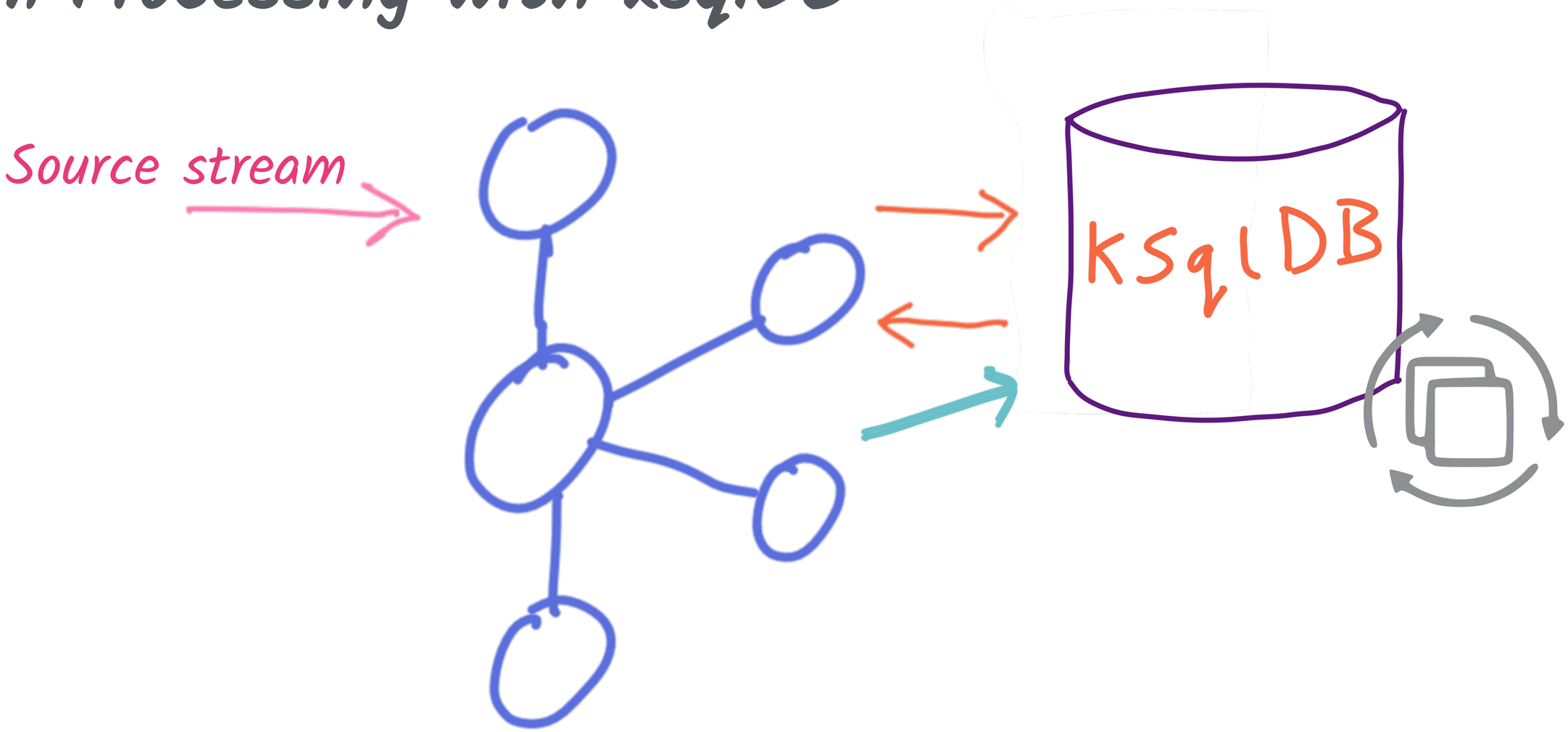
Stream Processing with ksqlDB



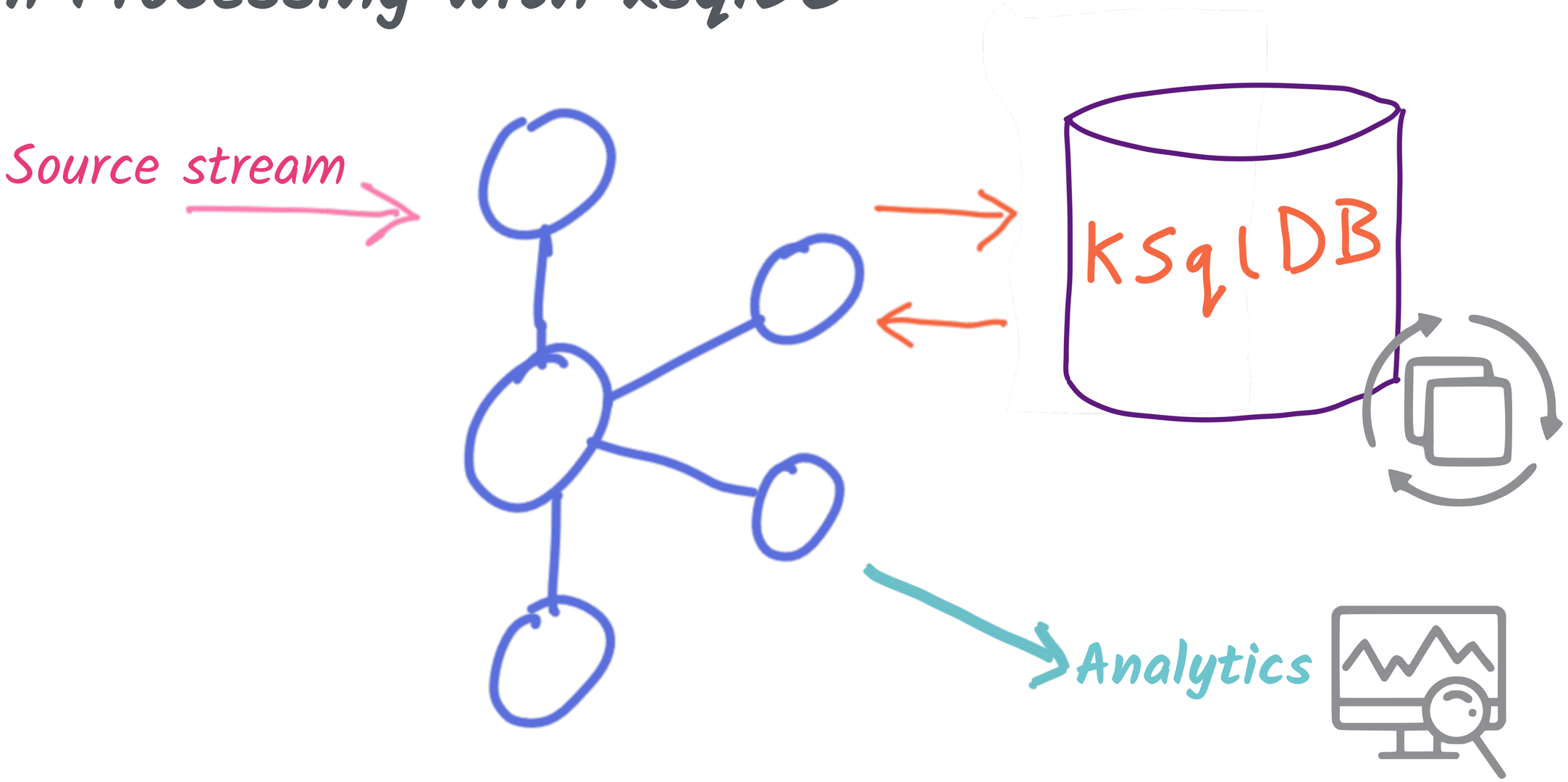
Stream Processing with ksqlDB



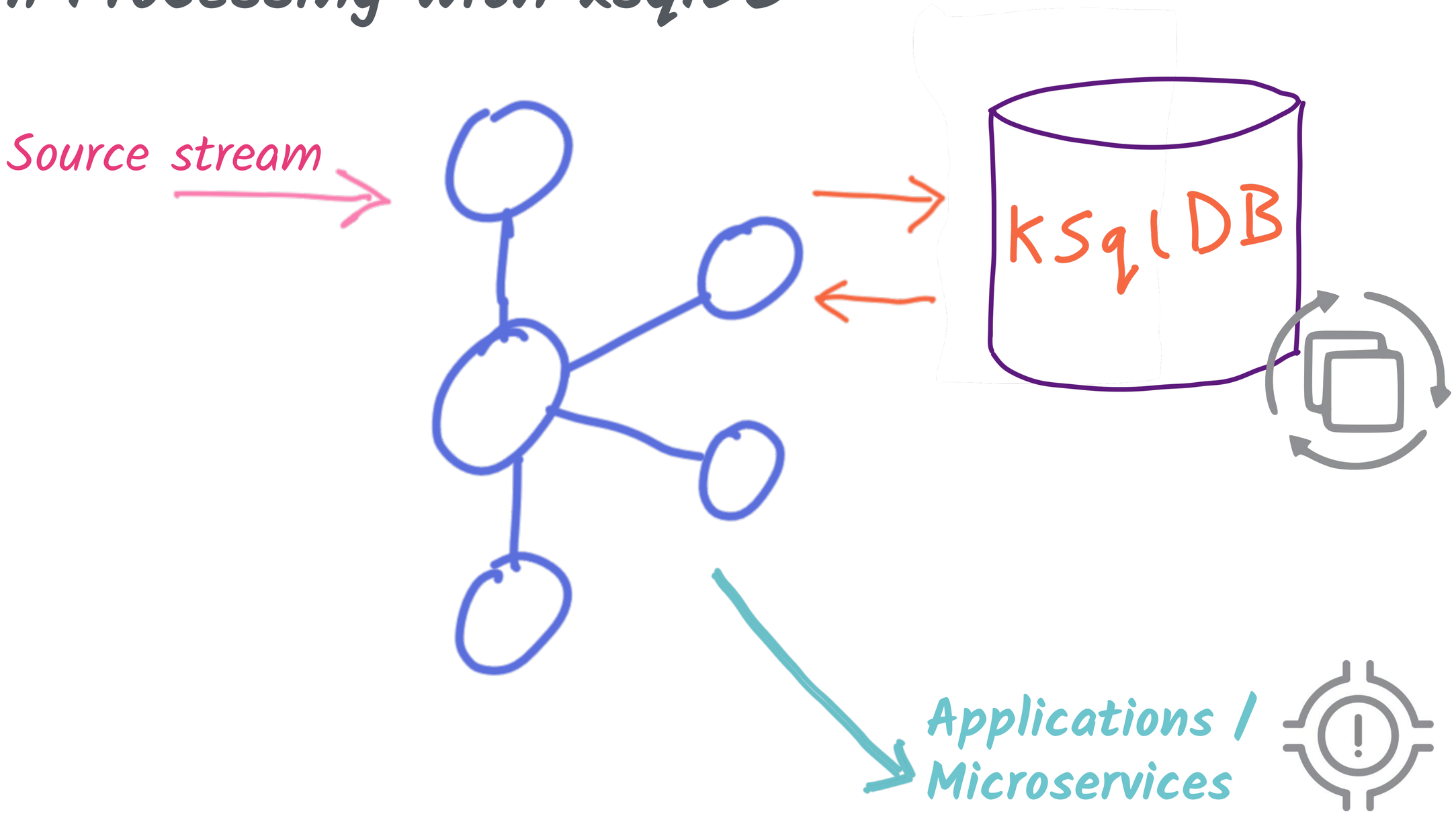
Stream Processing with ksqlDB



Stream Processing with ksqlDB

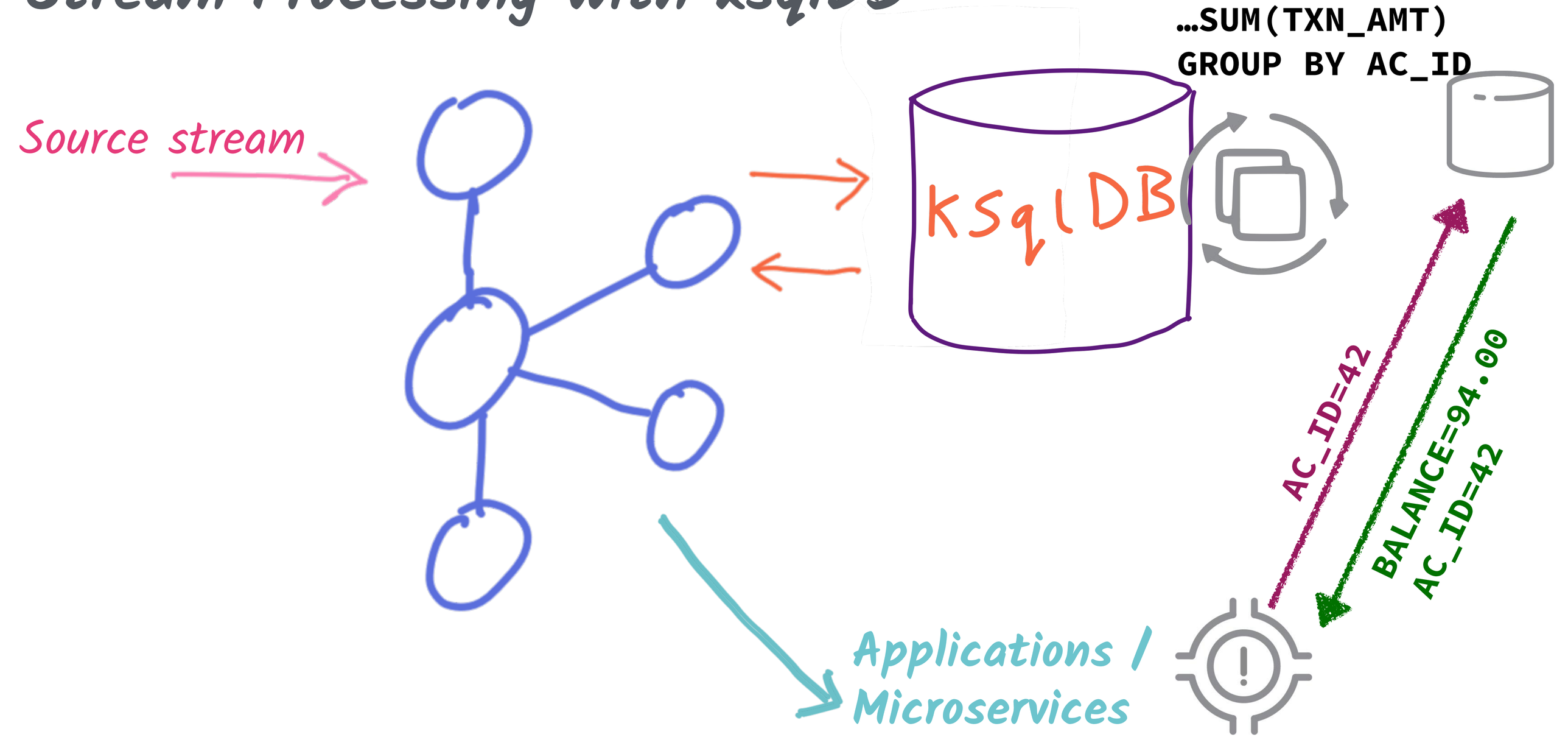


Stream Processing with ksqlDB

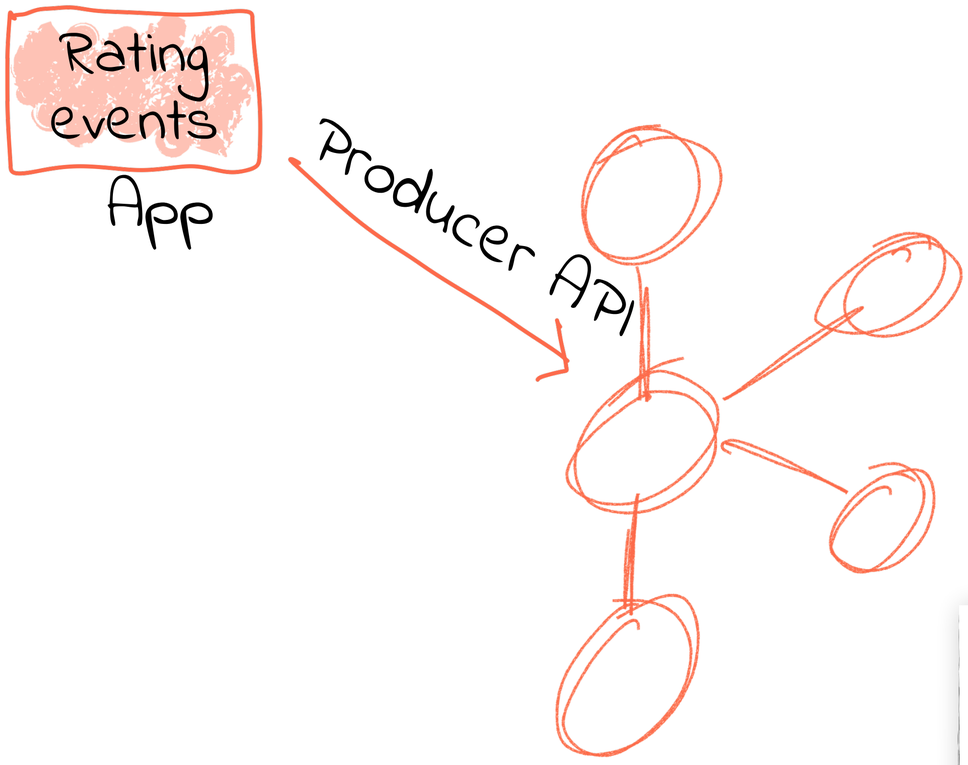


Stream Processing with ksqlDB

```
…SUM(TXN_AMT)
GROUP BY AC_ID
```

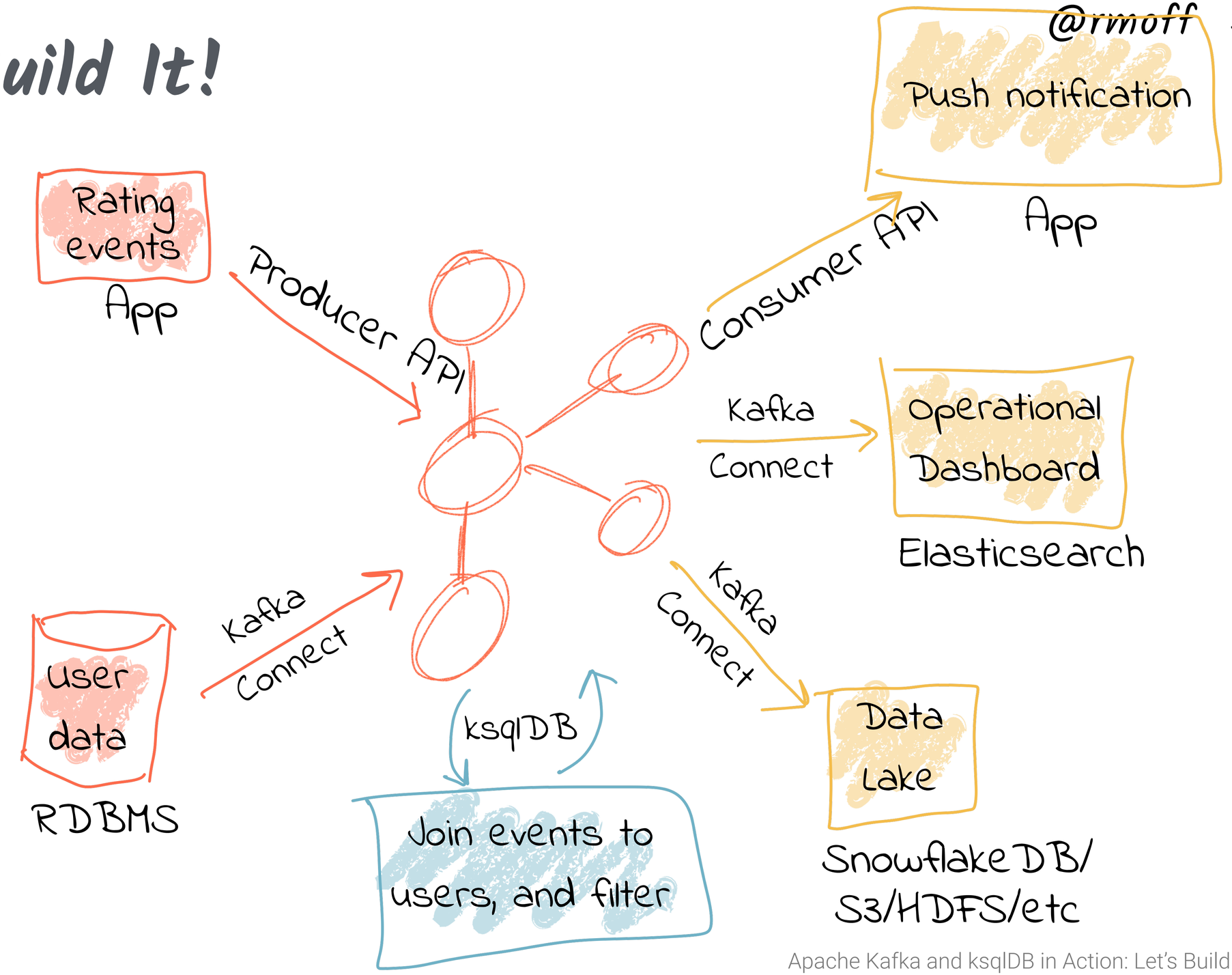


Let's Build It!

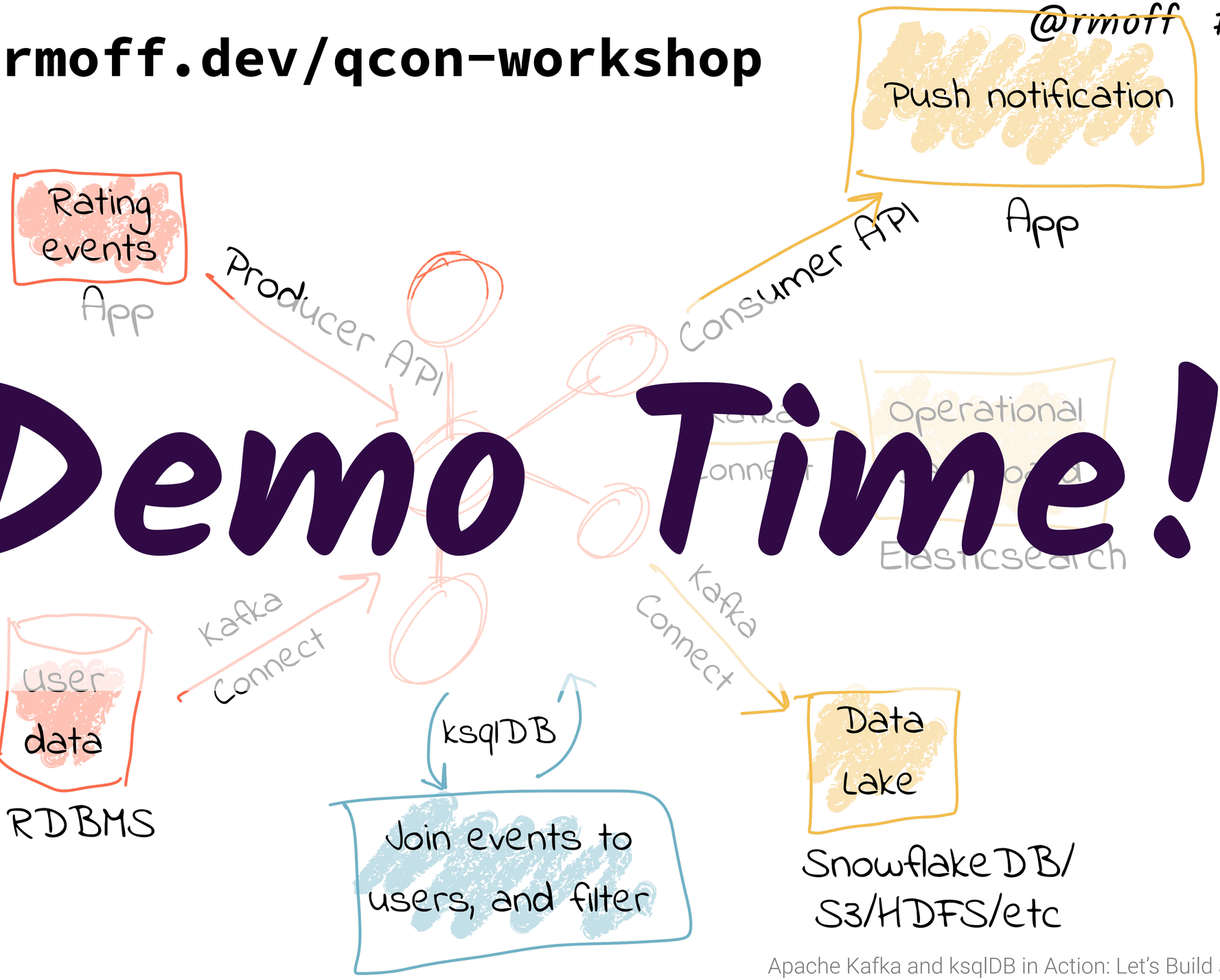


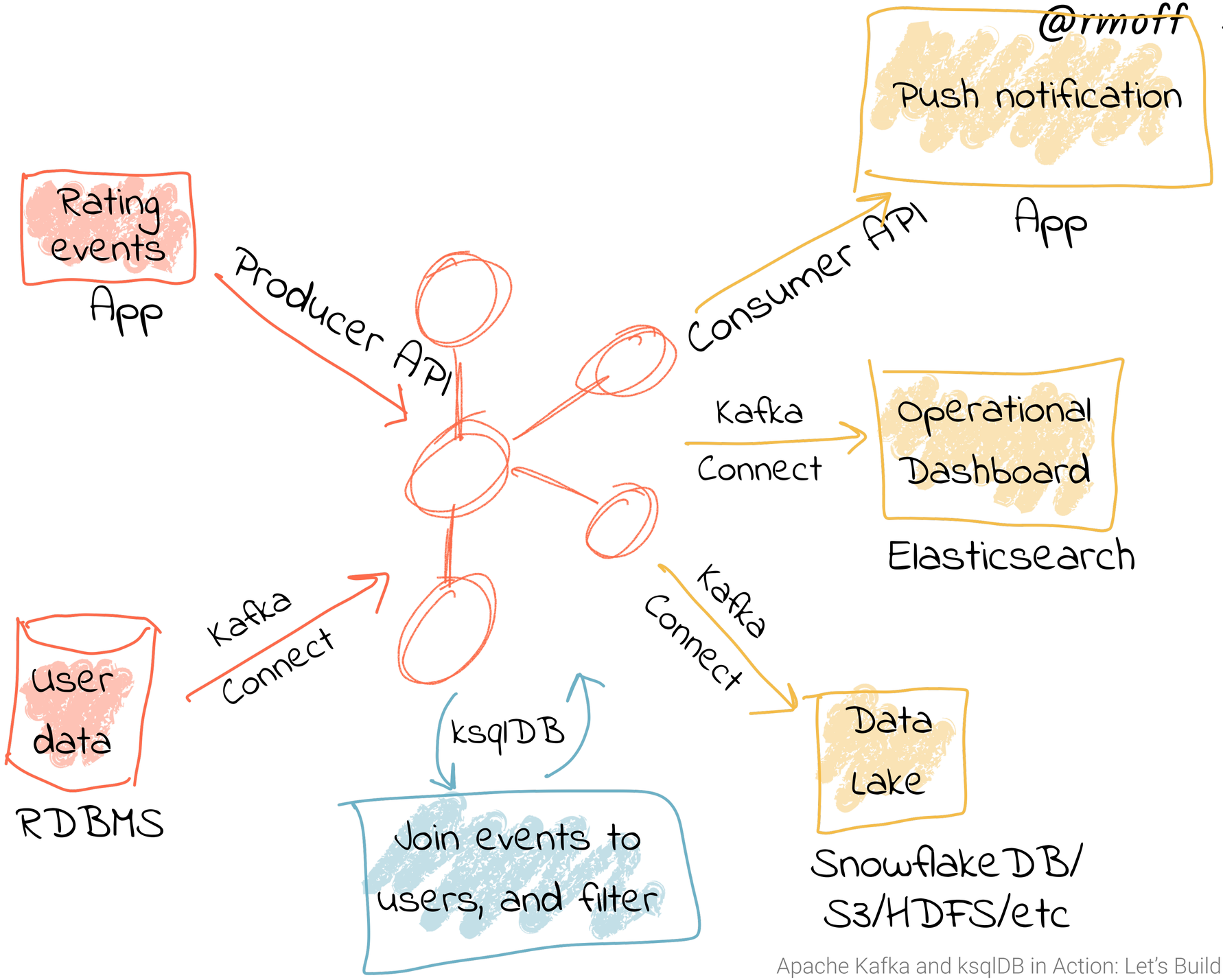

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

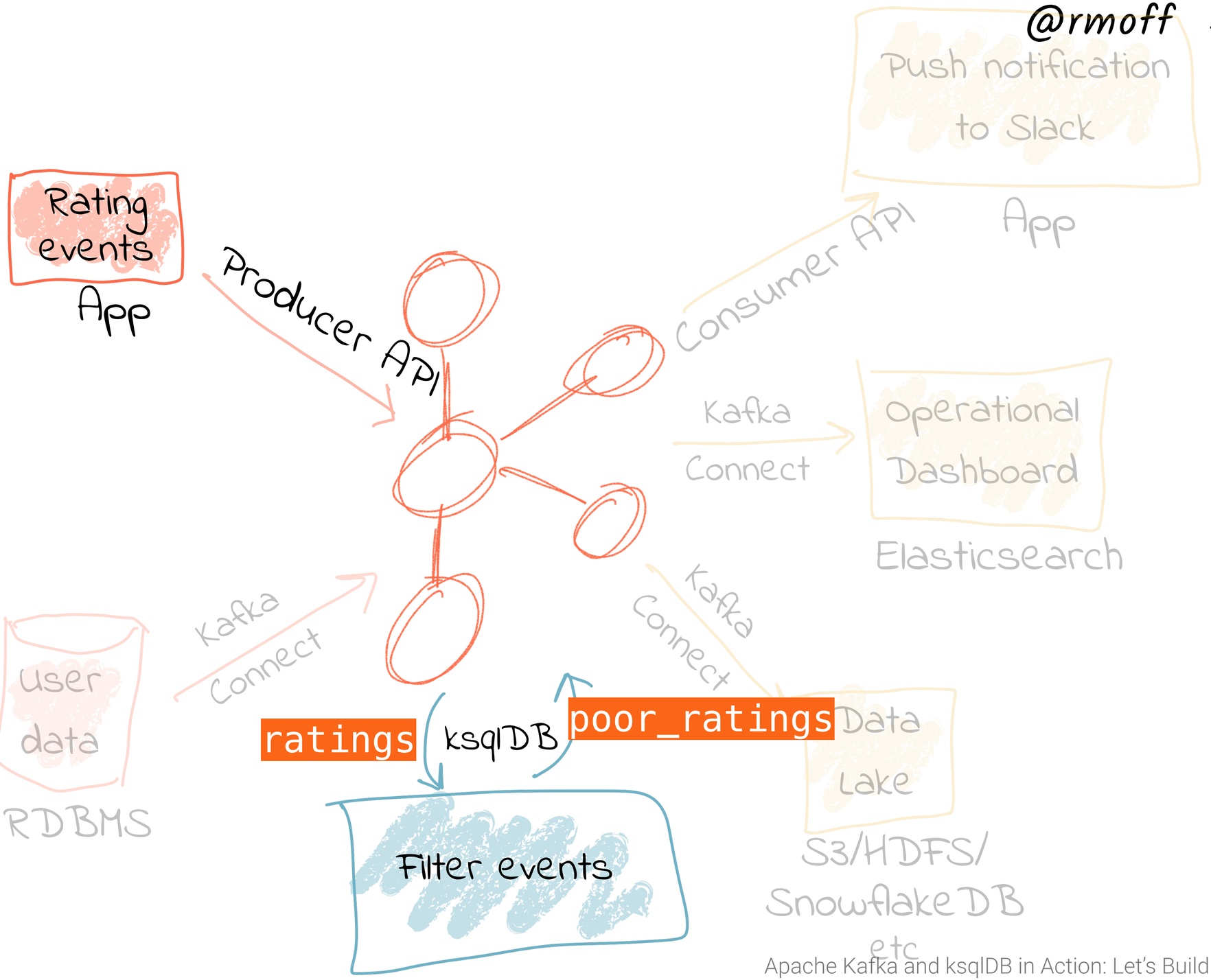
Let's Build It!



https://rmoff.dev/qcon-workshop

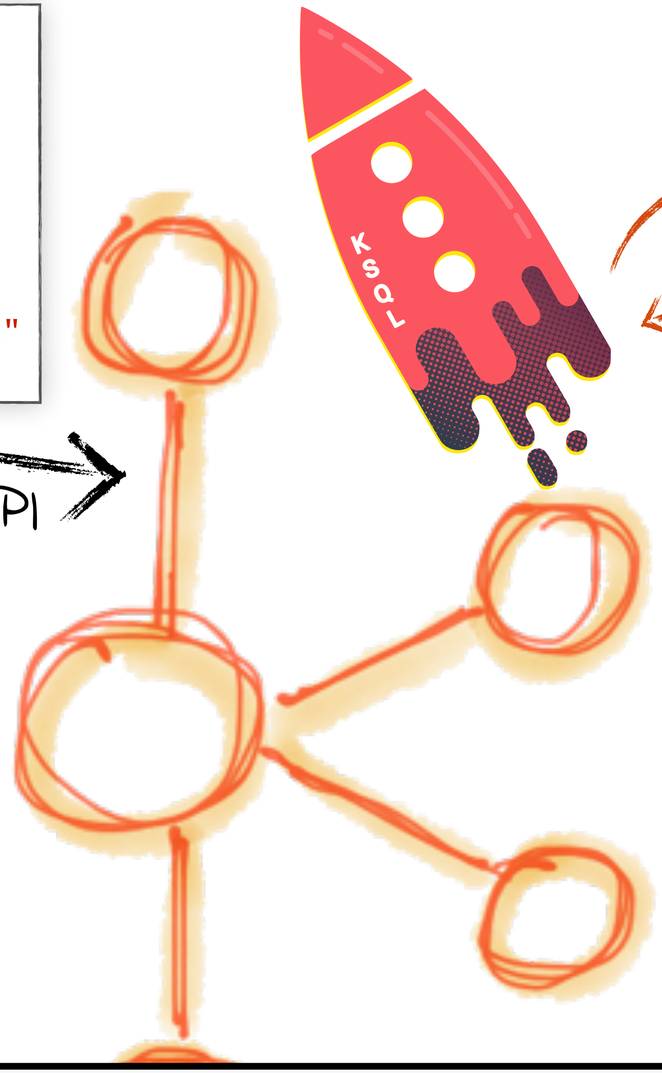



Push notification
App
Rating events
App
Producer API
Consumer API
Kafka Connect
Operational Dashboard
Elasticsearch
Kafka Connect
Kafka Connect
user data
RDBMS
ksqlDB
Join events to users, and filter
Data Lake
SnowflakeDB/ S3/HDFS/etc


Rating events
App
Producer API
Consumer API
Push notification to Slack
App
Kafka Connect
Operational Dashboard
Elasticsearch
Kafka Connect
User data
RDBMS
Kafka Connect
ratings
ksqlDB
poor_ratings
Data Lake
S3/HDFS/ SnowflakeDB etc
Filter events

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

Producer API

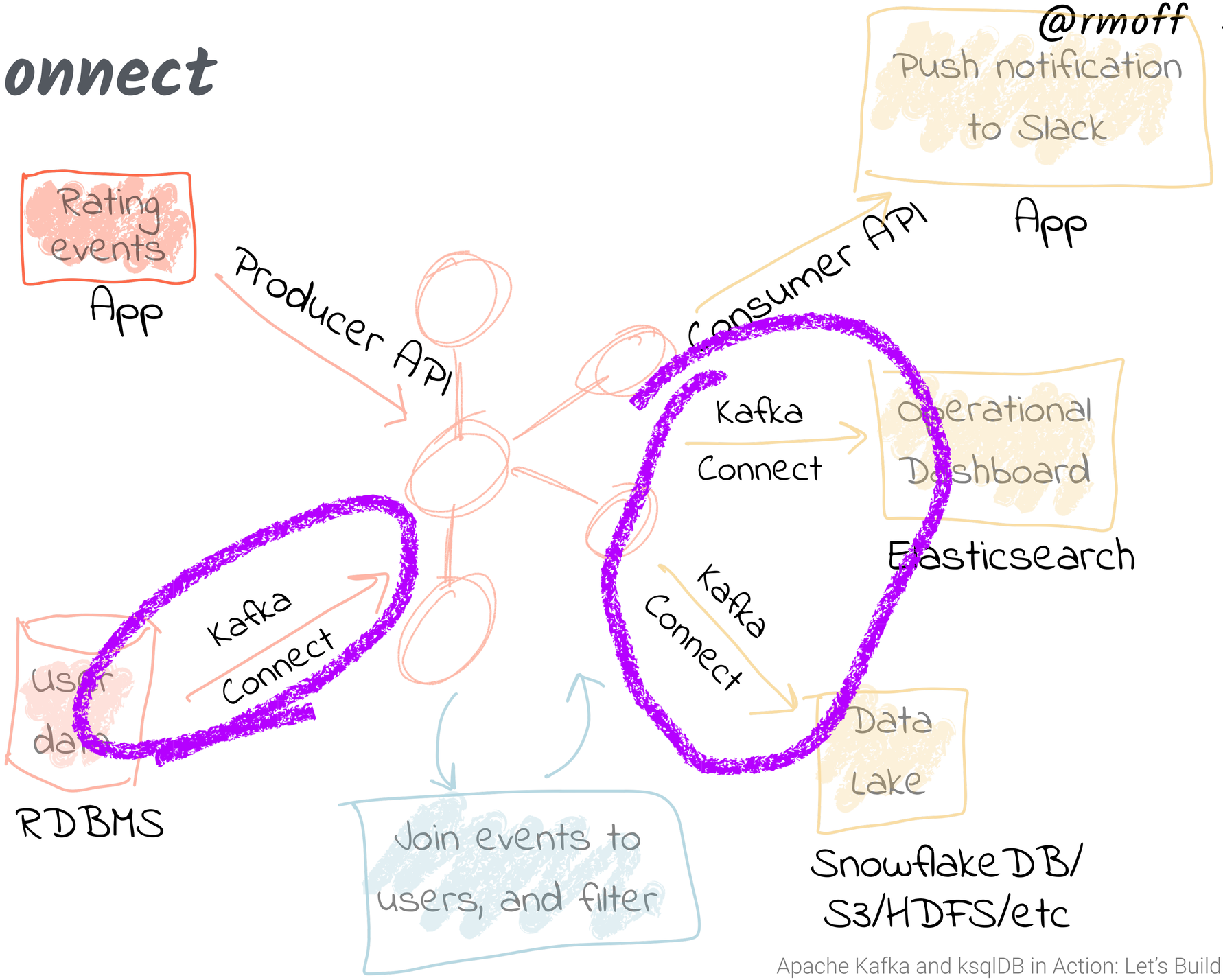


Filter all ratings where STARS<3

POOR_RATINGS

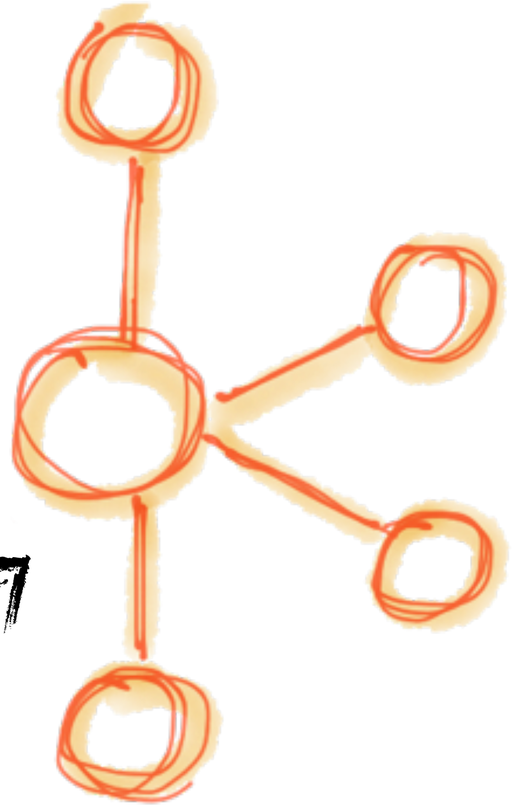
```
CREATE STREAM POOR_RATINGS AS
SELECT * FROM ratings WHERE STARS <3
```

Kafka Connect



```
{
  "rating_id": 2133,
  "user_id": 9,
  "stars": 1,
  "route_id": 7219,
  "rating_time": 1519402815063,
  "channel": "web",
  "message": "worst. flight. ever. #neveragain"
}
```


Producer API



Kafka Connect

Debezium

```
{
  "uid": 9,
  "name": "Neha",
  "locale": "en_US",
  "address_city": "Palo Alto",
  "elite": "P"
}
```

The Stream/Table Duality

Stream

Account ID	Amount
12345	+ £50
12345	+ £25
12345	- £60

Time

Table

Account ID	Balance
12345	£50

Account ID	Balance
12345	£75

Account ID	Balance
12345	£15

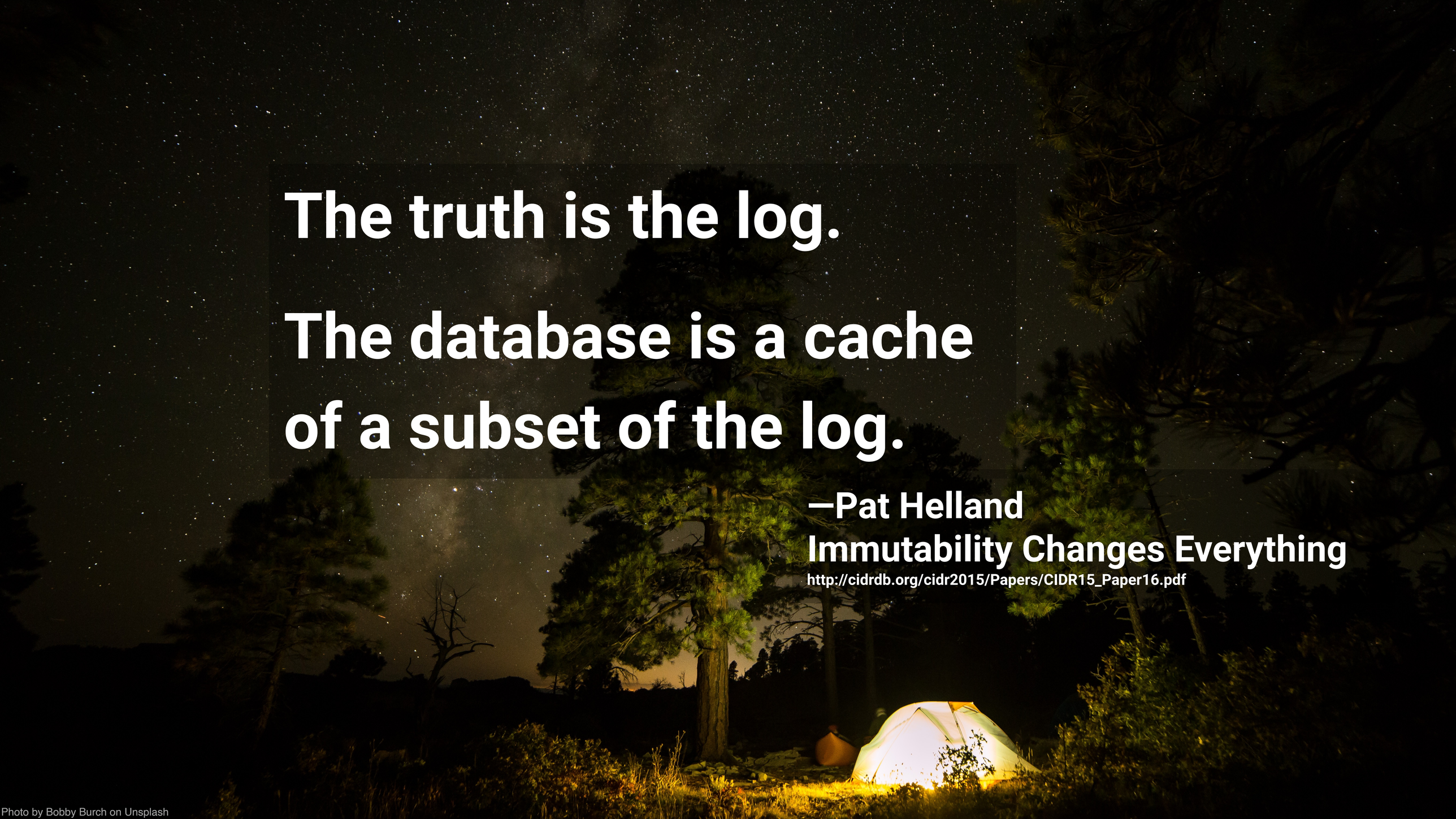
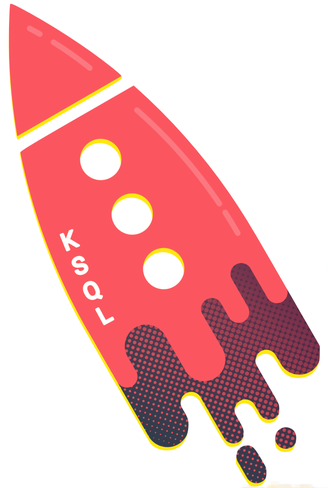


Photo by Bobby Burch on Unsplash

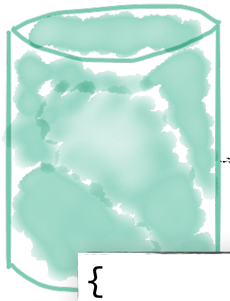
```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```


Producer API



Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA



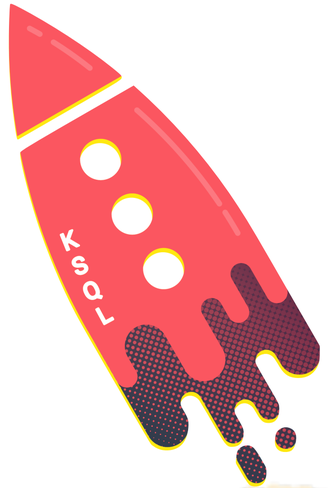
Kafka Connect


```
{
  "id": 3,
  "first_name
  "last_name"
  "email": "m
  "gender": "
  "club_statu
  "comments":
}
```


```
CREATE STREAM RATINGS_WITH_CUSTOMER_DATA AS
SELECT * FROM RATINGS
       LEFT JOIN CUSTOMERS ON R.ID=C.ID;
```

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

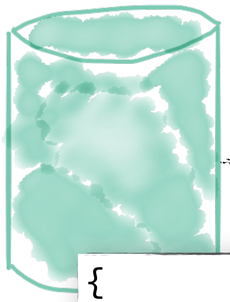

Producer API



Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA

Kafka Connect



Filter for just PLATINUM customers

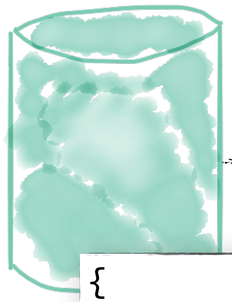
UNHAPPY_PLATINUM_CUSTOMERS


```
{
  "id": 3,
  "first_name
  "last_name"
  "email": "m
  "gender": "
  "club_statu
  "comments":
}
```


```
CREATE STREAM UNHAPPY_PLATINUM_CUSTOMERS AS
SELECT * FROM RATINGS_WITH_CUSTOMER_DATA
WHERE   STARS < 3
```

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304
 "channel": "web",
 "message": "worst. fli
}
```


```
CREATE TABLE RATINGS_BY_CLUB_STATUS AS
SELECT      CLUB_STATUS, COUNT(*)
FROM        RATINGS_WITH_CUSTOMER_DATA
            WINDOW TUMBLING (SIZE 1 MINUTES)
GROUP BY CLUB_STATUS;
```



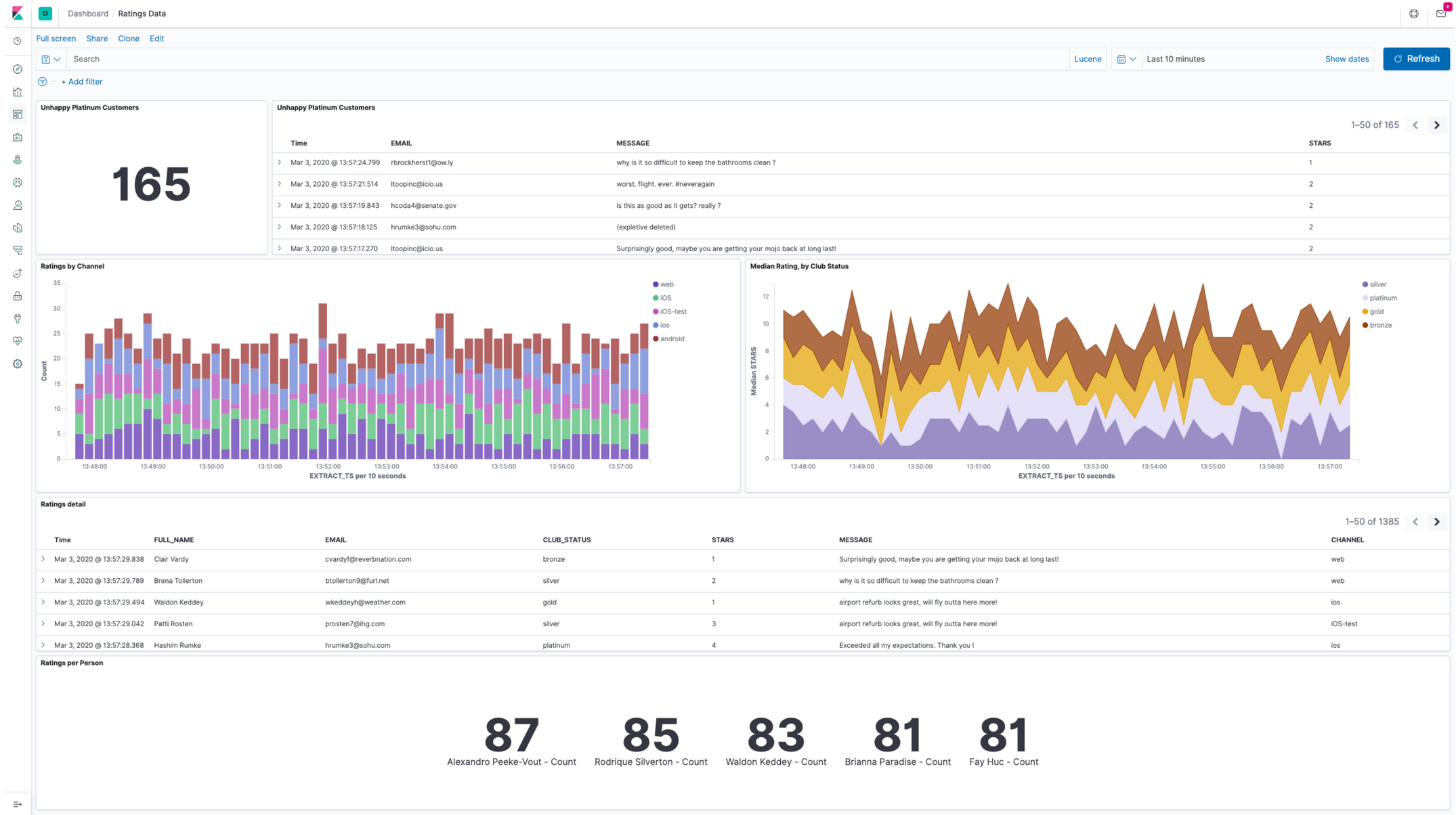
Kafka Connect

```
{
  "id": 3,
  "first_name": "Merilyn",
  "last_name": "Doughartie",
  "email": "mdoughartie1@dedecms.com",
  "gender": "Female",
  "club_status": "platinum",
  "comments": "none"
}
```


Aggregate per-minute by CLUB_STATUS

RATINGS_BY_CLUB_STATUS_1MIN

Kafka Connect → Elasticsearch



Want to learn more?

CTAs, not CATs
(sorry, not sorry)

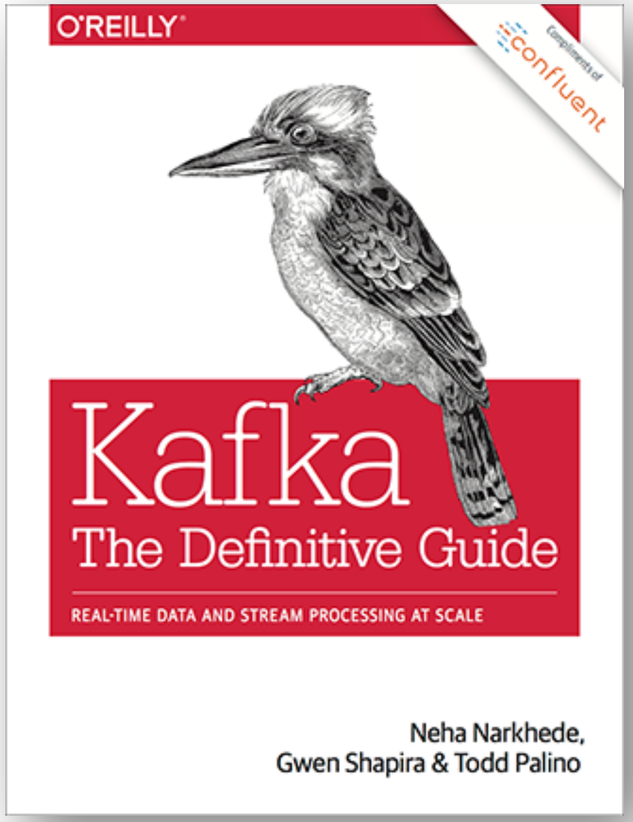
Learn Kafka.
Start building with
Apache Kafka at
Confluent Developer.
developer.confluent.io

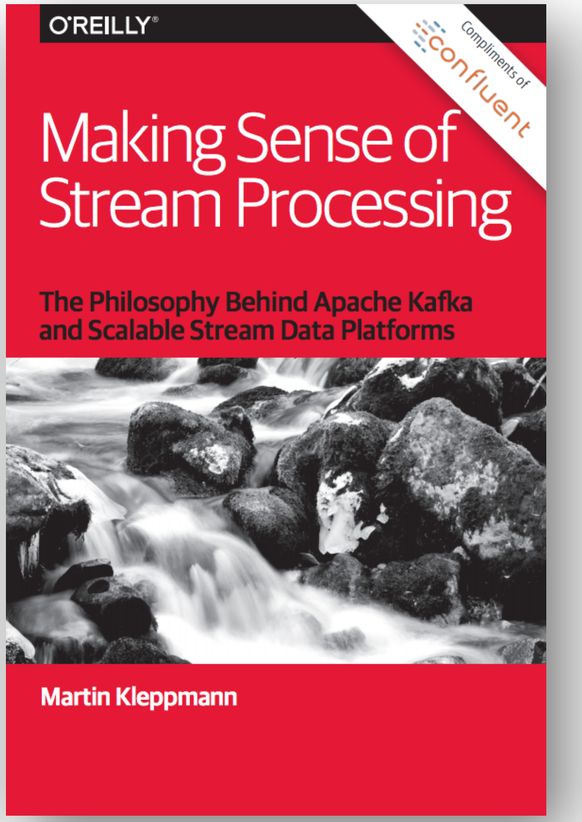
Confluent Community Slack group

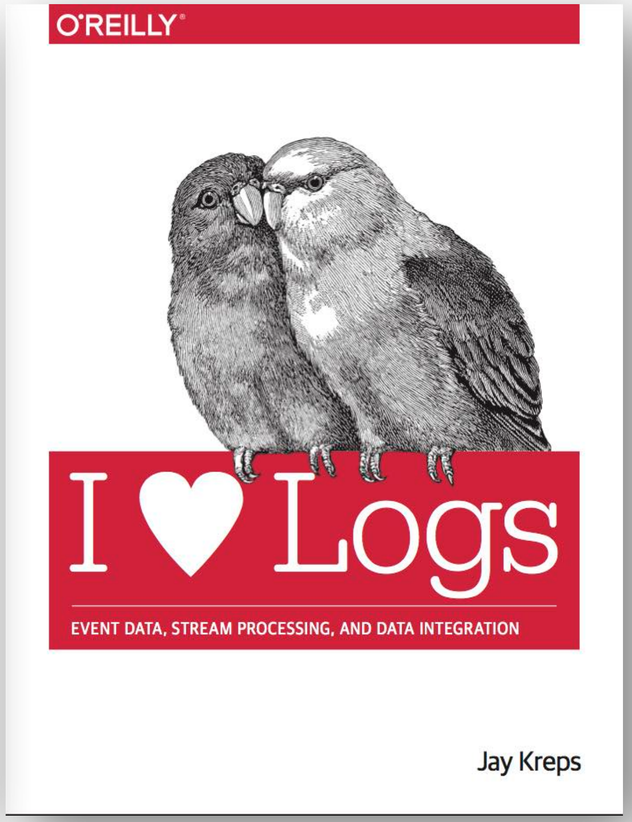
cnfl.io/slack

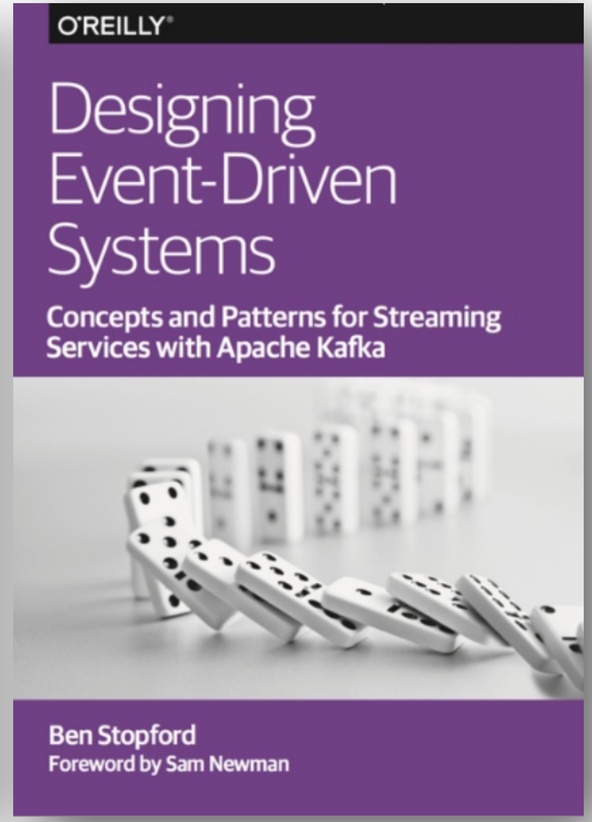
Free Books!

https://rmoff.dev/javabin









#EOF

Related Talks

- The Changing Face of ETL: Event-Driven Architectures for Data Engineers : https://rmoff.dev/oredev19-changing-face-of-etl
- From Zero to Hero with Kafka Connect: http://rmoff.dev/ksldn19-kafka-connect
- An introduction to ksqlDB: https://rmoff.dev/kssf19-ksql-slides
- No More Silos: Integrating Databases and Apache Kafka http://rmoff.dev/ksny19-no-more-silos